UNIVERSITY OF CALIFORNIA
Santa Barbara

Nonparametric Mixed-Effects Density Regression

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Statistics and Applied Probability

by

Chi-Yang Chiu

Committee in Charge:

Professor Yuedong Wang, Chair

Professor Wendy Meiring

Professor Roger Ingham

March 2015

The Dissertation of
Chi-Yang Chiu is approved:

Professor Wendy Meiring

Professor Roger Ingham

Professor Yuedong Wang, Committee Chairperson

March 2015

Nonparametric Mixed-Effects Density Regression

by

Chi-Yang Chiu

To my parents, Tsung-Ching Chiu and Yu-Chiao Chang, for their unwavering support, love, and encouragement.

Acknowledgements

First, I would like to express my most sincere gratitude to my advisor, Professor Yuedong Wang. With his patient guidance, profound knowledge and experiences, my research work is able to be carried out and grants me a solid foundation to thrive toward my future career path. I would like to thank my committee members, Professor Wendy Meiring and Professor Roger Ingham, for their unreserved advice, investing time and efforts to serve as my committee member.

It is also my pleasure to share the appreciation with my colleagues and friends at the department, Yuqi Chen, Yi-Tai Chiu, Fang-I Chu, Mark Dela, Jingyi He, Michael Nava, Jian Shi, Gaoyuan Tian, Xuwei Yang and Ling Zhu, have inspired me intellectually and supported emotionally through the graduate study. Our friendship and downtime at UCSB enriched my academic journey and will always be cherished.

I would also like to thank my siblings, Emily Chiu, Wen Chiu, and Justin Chiu. Their encouragement and blessing are my greatest assets. Lastly, my deepest gratitude goes to my wife, Nina Yang, and my daughter, Kelly Chiu. Their unconditional love has held me together through thick and thin. This dissertation work would not be possible without their emotional support.

Curriculum Vitæ
Chi-Yang Chiu

Education

2015	Doctor of Philosophy in Statistics and Applied Probability, Department of Statistics and Applied Probability, University of California, Santa Barbara.
2010	Master of Arts in Statistics, Department of Statistics and Applied Probability, University of California, Santa Barbara.
2005	Bachelor of Science in Mathematics, Tamkang University, New Taipei, Taiwan.

Experience

2009-2015	Teaching Assistant, Department of Statistics and Applied Probability, University of California, Santa Barbara.
2012	Biostatistics Intern, Department of Biostatistics, Allergan Medical, Santa Barbara.
2011	Teaching Associate, Department of Statistics and Applied Probability, University of California, Santa Barbara

Conference Presentations

2014	"Nonparametric Mixed-Effects Density Regression", with Dr. Yuedong Wang, Joint Statistical Meetings, Boston

Abstract

Nonparametric Mixed-Effects Density Regression

Chi-Yang Chiu

Conditional density provides the most informative summary of the relationship between independent and dependent variables. It enables us to examine the overall shapes of densities as well as summary characteristics such as quantiles and modes. Repeated measures designs are widely used in many areas such as agriculture, education and pharmaceutical sciences. The data from repeated measures designs are correlated. We develop a nonparametric method for conditional density estimation for repeated measures data. Specifically we propose nonparametric mixed-efffects density regression (NMDR) models. The NMDR models allow us to estimate conditional densities with fewer constraints on the form of densities when data are correlated. The models may be constructed using Smoothing Spline ANOVA (SS ANOVA) methods. Penalized marginal likelihood is used to estimate the density function as well as parameters. We use the stochastic approximation algorithm (SAA) with Newton-Raphson method for optimization, and Markov chain Monte Carlo (MCMC) for approximating integrals. An example from speech science is provided to illustrate the utility of our model.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Density Estimation

Density estimation is a procedure to estimate (or approximate) the underlying probability density function based on observed data. This is a fundamental problem in statistics. The density function f provides a description of the distribution of a random variable, which is important for prediction, inference, discrimination and classification.

One approach to density estimation is to assume that observations come from a known parametric family of distributions, for example the Normal distribution with mean μ and variance σ^2, or the exponential distribution with rate λ. In this situation, the density function is known except for a finite number of parameters. The parameters can be estimated by methods such as maximum likelihood, moments (Casella and Berger, 2002) or spacings (Ghosh and Jammalamadaka, 2001). This parametric approach is usually simple but sometimes the form of density is hard to specify.

In contrast to the parametric approach, the nonparametric method does not require a specific family for the density function. It lets the data speak for themselves and therefore is more flexible than the parametric approach. One of the most well known nonparametric approaches is the histogram, which is useful in data presentation and exploration. However, it is discontinuous and usually a smooth estimator is desirable. Many other nonparametric density estimation methods have been proposed, and we review some of these methods in Sections 1.2 to 1.4. We focus on spline based methods in the following subsections for different types of data.

1.2 Density Estimation for Independent and Identical Distributed Observations

The most basic type of data consists of independent and identically distributed (i.i.d.) observations. In this case, observations are a random sample, $Y_1, ..., Y_n$, from a certain distribution with density function $f(y)$. Our goal is to estimate the density function f.

Density estimation is complicated by two intrinsic constraints: nonnegativity constraint that $f \geq 0$ and the unity constraint that $\int_{\mathcal{Y}} f(y)dy = 1$ where $\mathcal{Y}$ is the domain. To deal with the nonnegativity constraint, O'Sullivan (1998), Stone (1990) and Kooperberg and Stone (1991) proposed to estimate the density function on the logarithm scale. They assumed that $\log f$ can be well approxi-

mated by a finite mixture of B-splines basis functions. This approach requires the specification of the number and locations of knots. Gu (1993) and Gu and Qiu (1993) proposed a smoothing spline approach. To deal with both the nonnegativity and the integrate to one constraints, they used the logistic transformation of the density $f = e^g / \int_{\mathcal{Y}} e^{g(y)} dy$ and estimated g through the minimization of penalized (negative) log likelihood. Details of the smoothing spline method will be discussed in Chapter 2.

In addition to the spline methods discussed above, Parzen (1962) developed a kernel method and Wahba (1981) used an orthogonal series such as Fourier series expansion to estimate f. Silverman (1986) provides an excellent introduction to nonparametric density estimation. Leonard (1978) and Lenk (1988, 1991) introduced and studied logistic Gaussian process priors for density estimation. Gehringer and Redner (1992) presented a nonparametric density estimate based on normalized tensor B-Splines. Efron and Tibshirani (1995) proposed a semiparametric technique by applying Poisson regression methods to specially designed parametric families through the kernel estimator. Dias (1998) proposed a hybrid spline approach which approximates the logistic transformed density g by a linear combination of B-spline basis functions, and estimated g by minimizing penalized (negative) log likelihood.

1.3 Conditional Density Estimation

Assessing the relationship between a dependent variable and one or more independent variables is of interest in many problems. For example, scientists may want to know how the distribution of blood pressure depends on gender. Regression analysis focuses on univariate characteristics such as conditional expectation, or quantiles of the dependent variable given the independent variables. The family of conditional distributions is usually assumed to be known, for example, as Gaussian or Poisson.

In some applications it is difficult, if not impossible, to specify a specific family of distributions, and the goal is to investigate covariate effects on the whole conditional density function. A conditional density estimate provides the most informative summary of the relationship between independent and dependent variables. It allows us to examine the overall shapes as well as summary characteristics such quantiles and modes.

Let (Y_i, X_i), $i = 1, ..., n$ be i.i.d. observations from a probability density $f(y, x)$ on a product domain $\mathcal{Y} \times \mathcal{X}$. The interest is to estimate the conditional density $f(y|x) = f(y, x)/\int_{\mathcal{Y}} f(y, x)dy$ of Y given X. Using the logistic transformation $f(y|x) = e^{g(y,x)}/\int_{\mathcal{Y}} e^{g(y,x)}dy$, Gu (1995) modeled the logistic conditional density $g(y, x)$ using tensor product smoothing splines. Details of this approach will be discussed in Chapter 2.

Other approaches include kernel methods proposed by Fan and Yim (2004) and Hall, Racine and Li (2004), orthogonal series methods proposed by Efromovich (2007), nonparametric Bayes methods proposed by Dunson, Pillai and Park (2007), the nonparametric empirical Bayes approach proposed by Dunson (2007), and semiparametric methods for comparing density differences in multi-sample situations studied by Qin and Zhang (2005) and Aubin and Leoni-Aubin (2008).

1.4 Conditional Density Estimation for Repeated Measures Data

Repeated measures data refers to data that include multiple measures from each subject. They arise in many areas such as agriculture, pharmacokinetics, epidemiology, medicine and social science. They are generated by observing each of a number of subjects repeatedly under varying conditions where the subjects are assumed to constitute a random sample from a population of interest.

Observations from the same subject are usually correlated and we are interested in estimating the density for the population as well as the density for each subject. In traditional regression analysis, mixed-effects models, including linear mixed-effects (LMEs), generalized linear mixed-effects (GLME) and nonlinear mixed-effects (NLME) models, are used for the analysis of repeated measures data. In such models, random effects are introduced in the conditional means to explain

the correlation caused by the subject effects. The application of smoothing spline methods in mixed-effects models has been studied by Wang (1998), Karcher and Wang (2001) and Ke and Wang (2001).

Traditional mixed-effects models assumed that data are from some specific family of distributions which sometimes is a strong assumption. The goal of nonparametric conditional density methods is to relax the assumption for the distribution of data. However, directly applying existing nonparametric conditional density methods developed for independent data to repeated measures data, will ignore the correlation within each subject. Hence we need nonparametric conditional density mixed effects models for repeated measures data.

One application of conditional density estimation for repeated measures data with flexible distributional assumptions was studied by Rodriguez, Dunson and Taylor (2009). In DNA repair studies, the measurements of interest are obtained from samples of cells from different individuals and the main interest focuses on understanding heterogeneity in the rates of DNA repair, adjusting for baseline damage and susceptibility to induced damage. In their study, Rodriguez, Dunson and Taylor used a finite mixture of Gaussian densities to approximate the unknown density. Even though this approach is more flexible than traditional mixed-effects models, the model proposed by Rodriguez, Dunson and Taylor (2009) is still a parametric model since the number of mixture components is finite and fixed.

Our research is motivated by an ongoing collaborative project with Professor Roger Ingham from the Department of Speech and Hearing Science at the

University of California - Santa Barbara, who is developing effective stuttering treatments. Dr. Ingham and colleagues had shown that a reduction in short phonated intervals (PIs) in the range of 30 to 150 ms is associated with decreased stuttering (Gow and Ingham 1992), and that purposefully reducing the number of short PIs resulted in the elimination of stuttering (Ingham, Kilgo, Ingham, Mogila, Belknap and Sanchez 2001).

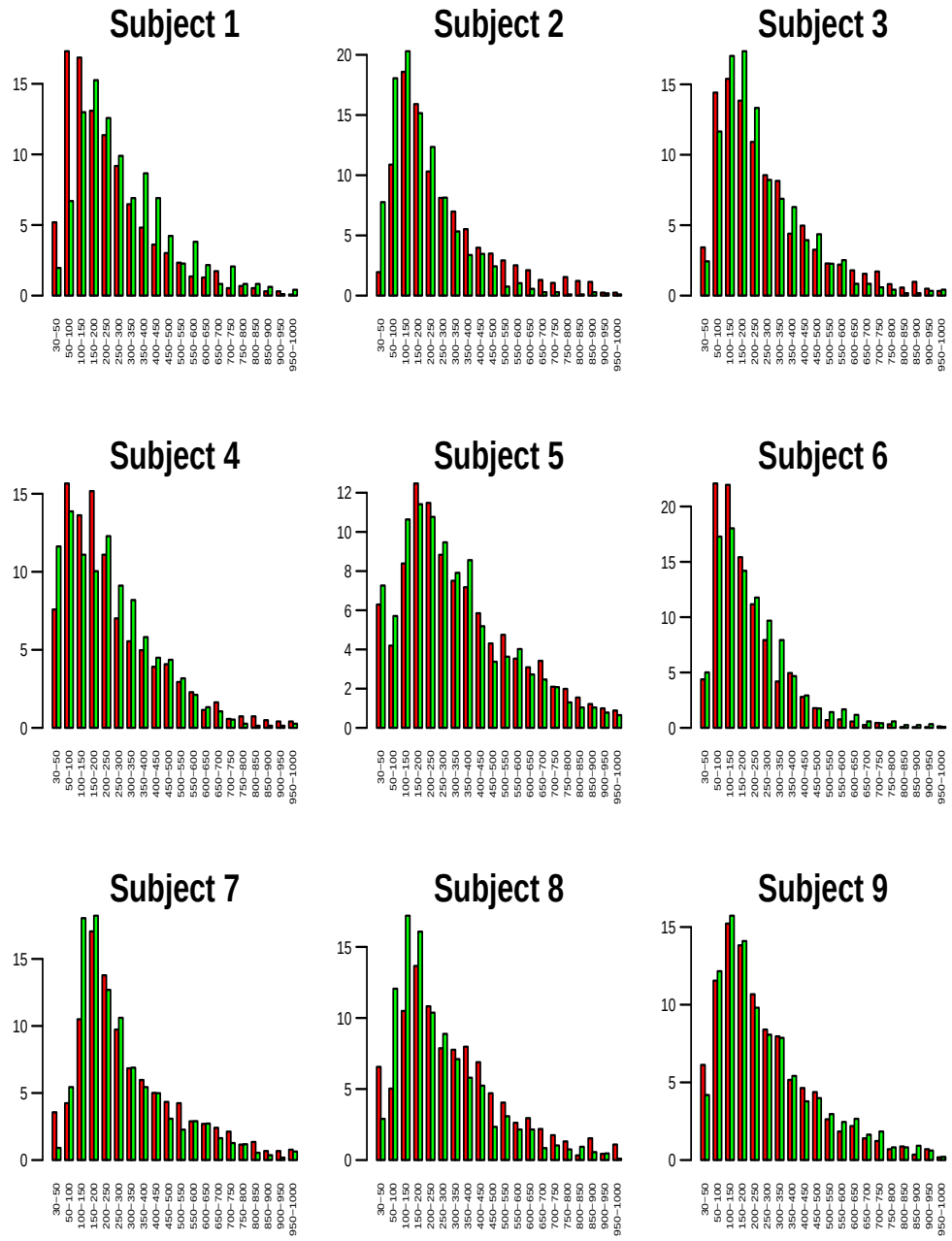
A PI is the elapsed time of a voiced unit of speech which is obtained by measuring the duration of vibration from the surface of the throat in between breaks of 10 ms or more. Observations are usually in the form of repeated measurements from multiple subjects under different conditions (e.g. rhythm, whispering, chorus and masking). Covariates may include gender and age. The goal is to compare density functions (especially in the short PI region) between speakers who stutter and normal subjects (or treatment and control) under different conditions.

The experiment involved 13 individuals who stuttered (11 of whom were males) and 13 control participants who were matched by age and gender. Subjects included both adults and adolsecent. Figure 1.1 displays the histograms of PIs for each subject. Figure 1.2 shows estimated density functions of PIs during oral reading for the 13 normal subjects and 13 people who stutter. The R package gss (Gu 2009) is used to estimate the density function for each subject separately. Visually, it appears that there is a large variation between subjects. Our goal is to compare density functions between people who stutter and normal subjects. The

dataset was provided by Professor Ingham. Additional details about the dataset and experiment can be seen in Godinho, Ingham, Davidow and Cotton (2006).

Several methods have been proposed to deal with correlated data. Hart and Vieu (1990) and Hall, Lahiri and Truong (1995) developed kernel density estimation methods for dependent data. Johnstone and Silverman (1997) proposed wavelet threshold estimators for data with correlated noise. Breunig (2001) proposed kernel density estimation methods for clustered data. Rodriguez and Horsty (2008) used nonparametric Bayesian approach to study dynamic density estimation for time-varying distributions. Rodriguez et al. (2009) used a finite mixture of Gaussian distributions to approximate the population density, and a hierarchical model for mixture weights, to assess heterogeneity across subjects as well as covariate effects. Griffin, Kolossiatis and Steel (2013) developed simultaneous Bayesian non-parametric modelling of several distributions. In this thesis, we will extend the SS ANOVA conditional density estimation method in Gu (1995) to the repeated measurement setting.

The dissertation is organized as follows. Chapter 2 reviews the smoothing spline density estimation method for independent data. Chapter 3 introduces our proposed model which extends smoothing spline density estimation to repeated measurement data. Chapter 4 describes the estimation procedures for our proposed model. In Chapter 5, we conduct extensive simulations to evaluate the performance of the proposed methods. Finally, Chapter 6 illustrates our proposed methods through the speech data.


Subject 1
Subject 2
Subject 3
Subject 4
Subject 5
Subject 6
Subject 7
Subject 8
Subject 9

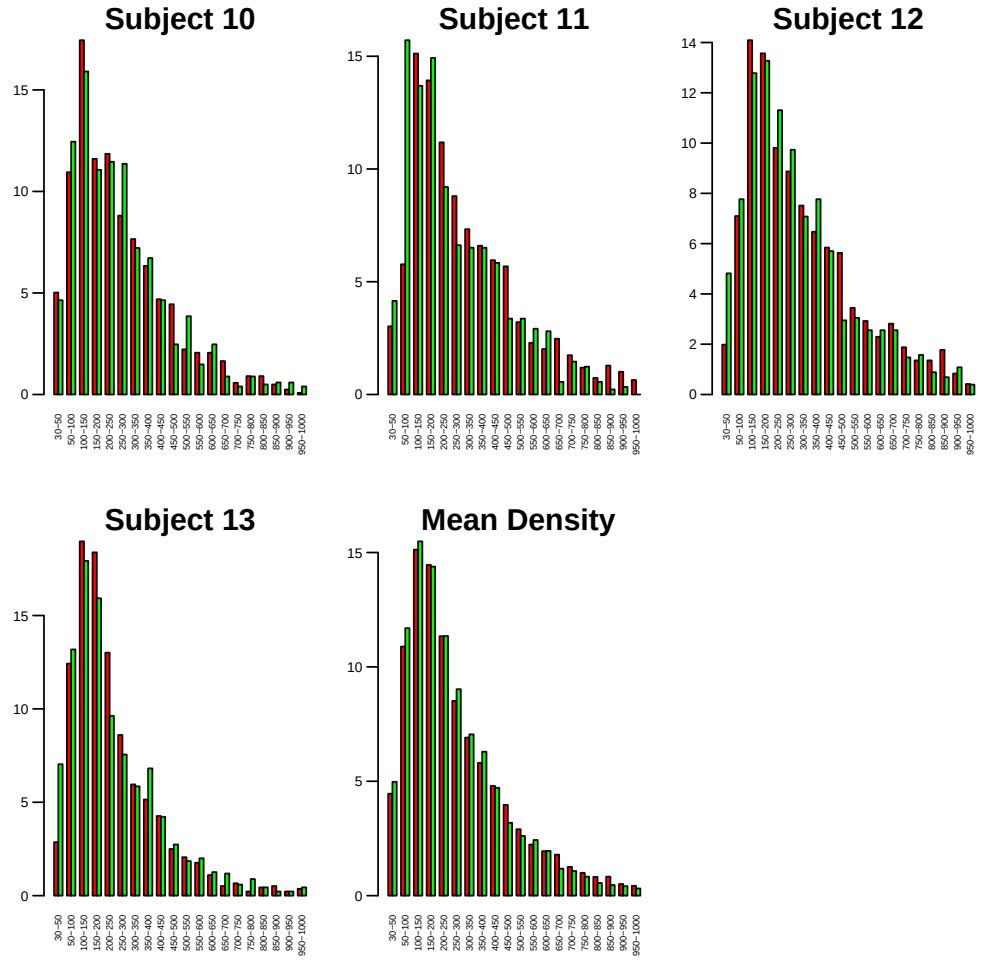


Figure 1.1: Histograms of PIs. In the i^{th} panel, for $i \in \{1.\ldots,13\}$, the red and green bins represent the PI proportion for the i^{th} matched pair of normal and stuttering subjects, respectively. The y-axis is bin proportion (in %), which is computed as the bin count divided by the total PI count across all time intervals for each subject. The x-axis is phonated interval (ms). The mean density shown in the final panel is computed by taking the average across all 13 subjects separately for each group.

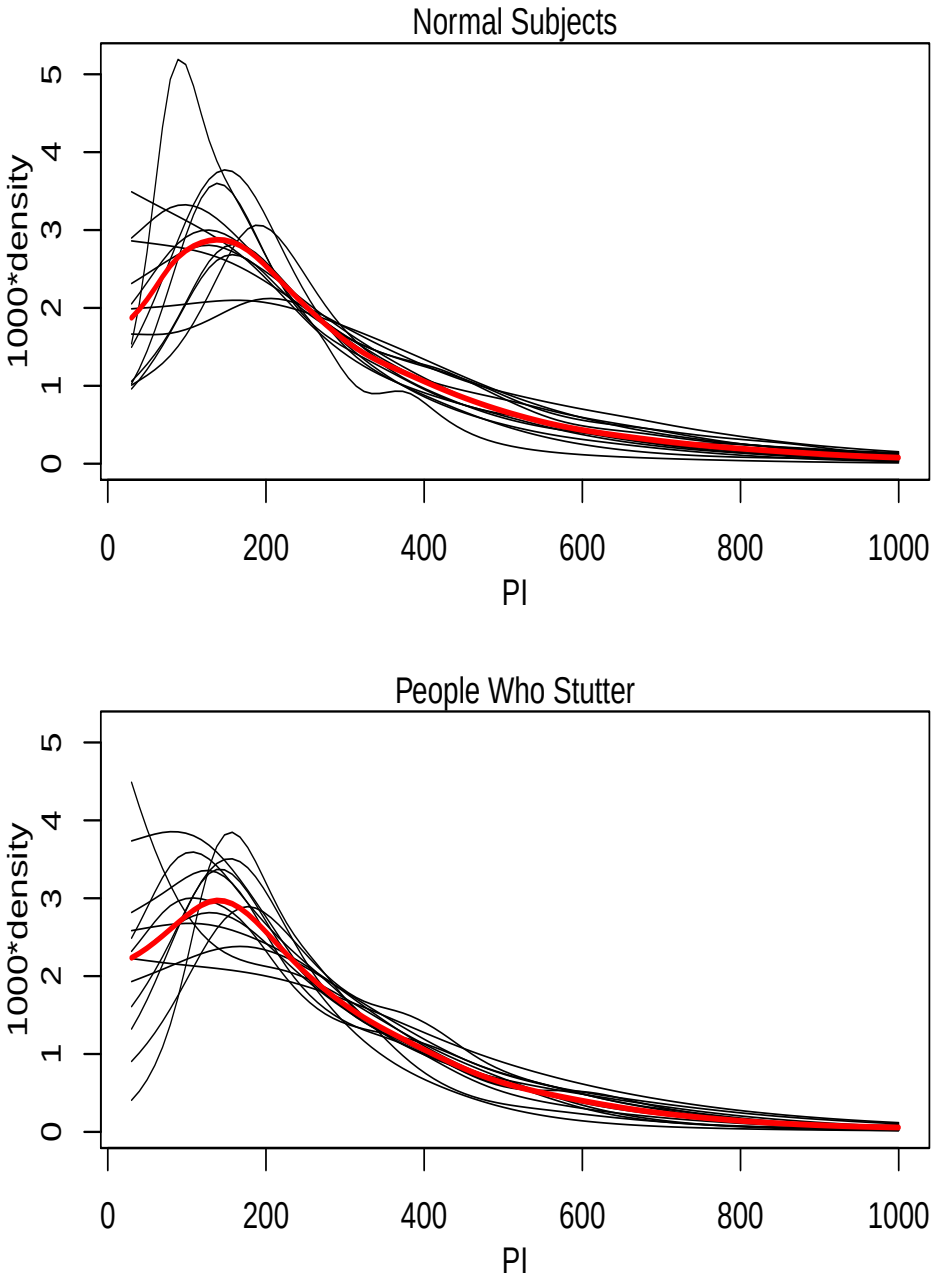


Figure 1.2: Estimated density functions of normal subjects (top) and people who stutter (bottom). Red thick line in each plot represents the mean of the estimated densities within each panel. Each mean density is calculated by taking the average across the 13 subject densities in the corresponding subject group.

Chapter 2

Smoothing Spline Density Estimation

2.1 Model Setting

Suppose we have observations $Y_i \overset{iid}{\sim} f(y)$, $y \in \mathcal{Y}$, where $\mathcal{Y}$ is an arbitrary set. In particular, the observation Y_i could be a scalar or a vector. Assume $f > 0$ on $\mathcal{Y}$. To be free of fundamental constraints, namely positivity and unity, the logistic density transform $f = e^g / \int_{\mathcal{Y}} e^{g(y)} dy$ will be employed. The goal here is to model and estimate the logistic density g. To ensure the logistic density transform is one-to-one, estimation must enforce a side condition on g such as $\int_{\mathcal{Y}} g(y)dy = 0$.

Denote $\mathcal{H}$ as the functional space of g with a side condition $\int_{\mathcal{Y}} g(y)dy = 0$. In smoothing spline density estimation, $\mathcal{H}$ is assumed to be a Reproducing Kernel Hilbert Space (RKHS) on $\mathcal{Y}$ with a reproducing kernel (RK) R. A RKHS is a Hilbert space in which every evaluation functional is continuous. The RK is a bivariate function on $\mathcal{Y}$, that is nonnegative definite and symmetric, $R(y_1, y_2) =$

$R(y_2, y_1)$. The RK has the reproducing property, $(R(y, \cdot), g(\cdot)) = g(y)$ where $(\cdot, \cdot)$ is the inner product in $\mathcal{H}$.

Applying a tensor sum decomposition, we decompose $\mathcal{H}$ into two subspaces

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1. \tag{2.1}$$

In (2.1) , $\mathcal{H}_0$ is a finite dimensional space with basis functions $\phi_1, ..., \phi_p$, and $\mathcal{H}_1$ is an RKHS with RK R_1. $\mathcal{H}_0$, often referred to as the null space, consists of functions that will not be penalized. For choosing $\mathcal{H}$, several factors including the domain $\mathcal{Y}$ and prior knowledge about the function g must be considered.

2.2 Penalized Likelihood Estimation

We use a penalized likelihood criterion to estimate g in (2.1). For the estimation of g, the likelihood is

$$\begin{aligned} L(g|Y_1, ..., Y_n) &= \prod_{i=1}^{n} f(Y_i) \\ &= \prod_{i=1}^{n} \frac{e^{g(Y_i)}}{\int_{\mathcal{Y}} e^{g(y)} dy}, \end{aligned}$$

and its logarithm

$$\log L = \sum_{i=1}^{n} \{g(Y_i) - \log \int_{\mathcal{Y}} e^{g(y)} dy\}.$$

Note, nature logarithms are used throughout this thesis. Denote P_1 as the orthogonal projection operator onto $\mathcal{H}_1$. We obtain the estimate of g via minimizing the following negative penalized log likelihood,

$$PL(g) = -\frac{1}{n} \sum_{i=1}^{n} \{g(Y_i) - \log \int_{\mathcal{Y}} e^{g(y)} dy\} + \frac{\lambda}{2} ||P_1 g||^2, \tag{2.2}$$

in $\mathcal{H}$. The negative log likelihood, $-\log L$, measures the goodness-of-fit of g to the data.

One should keep in mind that any member in $\mathcal{H}$ has to satisfy the predetermined side condition, $\int_{\mathcal{Y}} g(y)dy = 0$, so that the negative log likelihood is strictly convex. If the maximum likelihood estimate exists in the null space $\mathcal{H}_0$ which equips wtih the side condition $\int_{\mathcal{Y}} g(y)dy = 0$, the convexity of negative log likelihood establish the existence and uniqueness of the minimizer of (2.2) in $\mathcal{H}$. The proof for the convexity of negative log likelihood, and existence and uniqueness of the minimizer of (2.2) in $\mathcal{H}$ can be found in Gu (2013, Ch7).

The second term in PL(g) in (2.2) is a penalty term that penalizes the departure of our estimate of g from the null space $\mathcal{H}_0$. The smoothing parameter λ controls the trade-off between goodness-of-fit and departure from the null space $\mathcal{H}_0$. As λ goes to ∞, the limiting estimate falls in $\mathcal{H}_0$, which is a parametric model with a finite number of parameters. With $\lambda = 0$, one obtains the nonparametric maximum likelihood estimate, which is a sum of delta function spikes at the sample points, often referred to as the empirical distribution.

The solution to (2.2) might not fall in a finite dimensional space. Gu and Wang (2003) proposed to approximate the minimizer of (2.2) in a data-adaptive finite dimensional space

$$\mathcal{H}_q = \mathcal{H}_0 \oplus span\{R_1(Y_{i_j}, \cdot), j = 1, ..., q\}, \tag{2.3}$$

with $q \approx 10n^{2/9}$, where the set $\{Y_{i_1}, ..., Y_{i_q}\}$ is a random subset of $Y_1, ..., Y_n$. Set $\xi_j = R_1(Y_{i_j}, \cdot)$. By (2.3), any function g in $\mathcal{H}_q$ can be expressed as

$$g = \sum_{\nu=1}^{p} d_\nu \phi_\nu + \sum_{j=1}^{q} c_j \xi_j = \phi^T \mathbf{d} + \xi^T \mathbf{c}, \tag{2.4}$$

where $\boldsymbol{\phi} = (\phi_1, ..., \phi_p)^T$ and $\boldsymbol{\xi} = (\xi_1, ..., \xi_q)^T$ are vectors of functions and, $\mathbf{d} = (d_1, ..., d_p)^T$ and $\mathbf{c} = (c_1, ..., c_q)^T$ are vectors of coefficients. Then by substituting approximation (2.4) into (2.2), and noting that $||P_1 g||^2 = ||\sum_{j=1}^q c_j \xi_j||^2 = \sum_{j=1}^q \sum_{k=1}^q c_j c_k R_1(Y_{i_j}, Y_{i_k})$, for a fixed λ one can calculates the minimizer g_λ of (2.2) within the finite dimensional space $\mathcal{H}_q$. Specifically, for each fixed λ, estimator g_λ of g is found by minimizing

$$PL_\lambda(\mathbf{d}, \mathbf{c}) = -\frac{1}{n}\mathbf{1}^T(S\mathbf{d} + R\mathbf{c}) + \log \int_{\mathcal{Y}} \exp(\phi^T \mathbf{d} + \xi^T \mathbf{c}) dy + \frac{\lambda}{2}\mathbf{c}^T Q \mathbf{c}, \tag{2.5}$$

with respect to $\mathbf{d}$ and $\mathbf{c}$, where S is an $n \times p$ matrix with $(i, \nu)th$ element $\phi_\nu(Y_i)$, R is an $n \times q$ matrix with $(i, j)th$ element $\xi_j(Y_i) = R_1(\, Y_{i_j}, Y_i)$, and Q is a $q \times q$ matrix with (j, k)th element $\xi_j(Y_{i_k}) = R_1(Y_{i_j}, Y_{i_k})$.

The solution g_λ to (2.5) can be calculated using Newton iteration. Denote $\mu_g(h) = \int_{\mathcal{Y}} h(y) e^{g(y)} dy / \int_{\mathcal{Y}} e^{g(y)} dy$ and $V_g(h_1, h_2) = \mu_g(h_1 h_2) - \mu_g(h_1)\mu_g(h_2)$. Let $\widetilde{g} = \phi^T \widetilde{\mathbf{d}} + \xi^T \widetilde{\mathbf{c}} \in \mathcal{H}_q$. Take derivatives of (2.5) with respect to $\mathbf{d}$ and $\mathbf{c}$ at $\widetilde{g}$, the Newton updating equation is hence

$$\begin{pmatrix} V_{\boldsymbol{\phi},\boldsymbol{\phi}} & V_{\boldsymbol{\phi},\boldsymbol{\xi}} \\ V_{\boldsymbol{\xi},\boldsymbol{\phi}} & V_{\boldsymbol{\xi},\boldsymbol{\xi}} + \lambda Q \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} S^T \mathbf{1}/n - \mu_{\boldsymbol{\phi}} + V_{\boldsymbol{\phi},g} \\ R^T \mathbf{1}/n - \mu_{\boldsymbol{\xi}} + V_{\boldsymbol{\xi},g} \end{pmatrix}, \tag{2.6}$$

where $V_{\boldsymbol{\phi},\boldsymbol{\phi}}$ is an $p \times p$ matrix with (i, j)th element $V_{\widetilde{g}}(\phi_i, \phi_j)$, $V_{\boldsymbol{\phi},\boldsymbol{\xi}}$ is an $p \times q$ matrix with (i, j)th element $V_{\widetilde{g}}(\phi_i, \xi_j)$, $V_{\boldsymbol{\xi},\boldsymbol{\phi}}$ is the transpose of $V_{\boldsymbol{\phi},\boldsymbol{\xi}}$, $V_{\boldsymbol{\xi},\boldsymbol{\xi}}$ is an $q \times q$

matrix with (i,j)th element $V_{\widetilde{g}}(\xi_i, \xi_j)$, $\mu_{\boldsymbol{\phi}}$ is an p dimensional column vector with ith element $\mu_{\widetilde{g}}(\phi_i)$, $\mu_{\boldsymbol{\xi}}$ is an q dimensional column vector with ith element $\mu_{\widetilde{g}}(\xi_i)$, $V_{\boldsymbol{\phi},g}$ is an p dimensional column vector with ith element $V_{\widetilde{g}}(\phi_i, \widetilde{g})$ and $V_{\xi,g}$ is an q dimensional column vector with ith element $V_{\widetilde{g}}(\xi_i, \widetilde{g})$.

One simple example is to consider $\mathcal{Y} = [0,1]$ and the functional space $\mathcal{H}$ to be a Sobolev space W_2^m defined as follows:

$$W_2^m = \{g : g^{(i)} \text{ are absolutely continuous}, i = 1, ..., m-1, \int_0^1 [g^{(m)}(y)]^2 dy < \infty\},$$

where $g^{(j)}$ is the jth derivative of $g(y)$ with respect to y. When $m = 2$ then $||P_1 g||^2 = \int_0^1 [g^{(2)}(y)]^2 dy$ and with side condition $\int_0^1 g(y)dy = 0$, (2.3) has $\mathcal{H}_0 = \{y - 0.5\}$ and $R_1(Y_{i_j}, \cdot) = k_2(Y_{i_j})k_2(\cdot) - k_4(|Y_{i_j} - \cdot|)$, where k_2, k_4 are scaled Bernoulli polynomials. Hence, the estimate of g in (2.4) can be represented as

$$g_\lambda(y) = d_\lambda \times (y - 0.5) + \sum_{j=1}^{q} c_{j,\lambda}\xi_j,$$

where $\xi_j(y) = k_2(Y_{i_j})k_2(y) - k_4(|Y_{i_j} - y|)$. Note that the first four scaled Bernoulli polynomials are

$$\begin{aligned}
k_0(x) &= 1, \\
k_1(x) &= x - 0.5, \\
k_2(x) &= \frac{1}{2}\{k_1^2(x) - \frac{1}{12}\}, \\
k_4(x) &= \{k_1^4(x) - \frac{1}{2}k_1^2(x) + \frac{7}{240}\}.
\end{aligned}$$

2.3 Smoothing Parameter Selection

To estimate λ, one needs a measure for closeness of the estimated density $f_\lambda = e^{g_\lambda}/\int_{\mathcal{Y}} e^{g_\lambda(y)}dy$ to the true density $f = e^g/\int_{\mathcal{Y}} e^{g(y)}dy$. One choice is the Kullback-Leibler (K-L) loss,

$$\begin{aligned} KL(g, g_\lambda) &= E_f[\log(f/f_\lambda)] \\ &= \mu_g(g - g_\lambda) - \log \int_{\mathcal{Y}} e^{g(y)}dy + \log \int_{\mathcal{Y}} e^{g_\lambda(y)}dy, \end{aligned}$$

where $\mu_g(h) = \int_{\mathcal{Y}} h(y)e^{g(y)}dy/\int_{\mathcal{Y}} e^{g(y)}dy$ is defined in Section 2.2. An optimal λ can be considered as the one that minimizes $KL(g, g_\lambda)$. Dropping the terms in $KL(g, g_\lambda)$ that are independent of g_λ, one has the relative K-L loss,

$$RKL(g, g_\lambda) = \log \int_{\mathcal{Y}} e^{g_\lambda(y)}dy - \mu_g(g_\lambda). \tag{2.7}$$

The second term $\mu_g(g_\lambda)$ in (2.7) depends on the unknown density, which is needed to be estimated. A naive way to estimate $\mu_g(g_\lambda)$ is to use the sample mean $n^{-1}\sum_{i=1}^n g_\lambda(Y_i)$, but the resulting estimate of $RKL(g, g_\lambda)$ would simply be the minus log likelihood which leads to $\lambda = 0$. The naive estimate, $n^{-1}\sum_{i=1}^n g_\lambda(Y_i)$, also leads to a biased estimate of $RKL(g, g_\lambda)$ since the same samples $Y_i's$ are used to obtain and assess the estimate g_λ. To remedy this problem, Gu and Wang (2003) use standard cross-validation method to estimate $\mu_g(g_\lambda)$ by $\widetilde{\mu}_g(g_\lambda) = n^{-1}\sum_{i=1}^n g_\lambda^{[i]}(Y_i)$ where $g_\lambda^{[i]}$ is the minimizer of the delete-one version of (2.2),

$$-\frac{1}{n-1}\sum_{j\neq i}^{n}\{g(Y_j) - \log \int_{\mathcal{Y}} e^{g(y)}dy\} + \frac{\lambda}{2}||P_1 g||^2. \tag{2.8}$$

The delete-one estimates $g_\lambda^{[i]}$ are expensive to compute. To reduce computation, one may find the quadratic approximation of (2.2) at $\widetilde{g} = g_\lambda$ by applying the second-order Taylor expansion on $\log \int_{\mathcal{Y}} e^{g(y)} dy$, and then compute the minimizer $g_{\lambda,\widetilde{g}}^{[i]}$ of the delete-one-version of that.

Estimating $\mu_g(g_\lambda)$ in (2.7) by $n^{-1} \sum_{i=1}^n g_{\lambda,\widetilde{g}}^{[i]}(Y_i)$ leads to the delete-one cross-validation estimate of $RKL(g, g_\lambda)$,

$$CV(\lambda) = -\frac{1}{n} \sum_{i=1}^{n} \{g_\lambda(Y_i) - \log \int_{\mathcal{Y}} e^{g_\lambda(y)} dy\} + \frac{tr(P_{\mathbf{1}}^{\perp} \breve{R}^T H^{-1} \breve{R} P_{\mathbf{1}}^{\perp})}{n(n-1)}, \tag{2.9}$$

where H is the left-hand-side matrix in (2.6), $\breve{R} = (S, R)$, and $P_{\mathbf{1}}^{\perp} = I - \mathbf{1}\mathbf{1}^T/n$ is an $n \times n$ matrix.

To prevent occasional under-smoothing, Gu (2013, Ch7) suggests the following modified CV score,

$$CV_\alpha(\lambda) = -\frac{1}{n} \sum_{i=1}^{n} \{g_\lambda(Y_i) - \log \int_{\mathcal{Y}} e^{g_\lambda(y)} dy\} + \alpha \frac{tr(P_{\mathbf{1}}^{\perp} \breve{R}^T H^{-1} \breve{R} P_{\mathbf{1}}^{\perp})}{n(n-1)}, \tag{2.10}$$

which is obtained by simply multiplying the trace term in (2.9) by a constant $\alpha > 1$. Details about the optimal α value and smoothing parameter selection for density estimation can be found in Gu (2013, Ch7).

2.4 Conditional Density Estimation

2.4.1 Introduction

Let $(Y_i, X_i), i = 1, ...n$ be *i.i.d.* observations from a probability density $f(y, x)$ on a product domain $\mathcal{Y} \times \mathcal{X}$. We are interesting in estimating the conditional

density $f(y|x) = f(y,x)/\int_{\mathcal{Y}} f(y,x)dy$ of Y given X, without assuming any form of parametric model for $f(y,x)$ or $f(y|x)$. Gu (1995) extended the development in smoothing spline density estimation to the conditional density estimation on general domain.

The formulation is similar to the smoothing spline density estimation. The logistic transform $f(y|x) = e^{g(y,x)}/\int_{\mathcal{Y}} e^{g(y,x)}dy$ is employed, enabling one to estimate g instead of f to naturally impose the positivity and unity constraints. Certain side conditions are also needed to make the transform one-to-one. The choice for side conditions will be briefly mentioned later in Section 2.4.2. The bivariate function g is defined on the product domain $\mathcal{Y}\times\mathcal{X}$. To model the joint function g, one may use smoothing spline ANOVA (SS ANOVA) decomposition of the tensor product RKHS which will be introduced in the following subsection.

2.4.2 Tensor Product RKHS

Denote $\mathcal{H}$ as the functional space for the joint function g and consider it to be a tensor product RKHS on $\mathcal{Y}\times\mathcal{X}$. By applying a tensor sum decomposition, we have

$$\mathcal{H} \triangleq \mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)}, \tag{2.11}$$

where marginal spaces $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ are RKHS's on $\mathcal{Y}$ and $\mathcal{X}$ respectively.

Consider averaging operators $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ on $\mathcal{H}^{(1)}$and $\mathcal{H}^{(2)}$ respectively, and denote I as the identity map. The joint function g can be decomposed into

$$\begin{aligned} g &= \{\mathcal{A}^{(1)} + (I - \mathcal{A}^{(1)})\}\{\mathcal{A}^{(2)} + (I - \mathcal{A}^{(2)})\}g \\ &= \{\mathcal{A}^{(1)}\mathcal{A}^{(2)} + \mathcal{A}^{(1)}(I - \mathcal{A}^{(2)}) + (I - \mathcal{A}^{(1)})\mathcal{A}^{(2)} + (I - \mathcal{A}^{(1)})(I - \mathcal{A}^{(2)})\}g \\ &\triangleq \mu + g_x(x) + g_y(y) + g_{yx}(y, x), \end{aligned} \tag{2.12}$$

where μ is constant, $g_y(y)$ and $g_x(x)$ are the main effects, and $g_{yx}(y, x)$ is the interaction. The choice of operators $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ depends on the domain of the marginal functions of y and x. For example, let $\mathcal{Y} = [a, b]$ and $\mathcal{X} = \{1, ..., m\}$, one may consider $\mathcal{A}^{(1)}$ and $\mathcal{A}^{(2)}$ such that $\mathcal{A}^{(1)}g = \int_a^b g(y, x)dy/(b - a)$ and $\mathcal{A}^{(2)}g = \sum_{x=1}^m g(y, x)/m$.

The decomposition of the bivariate function g in (2.12) is called the two-way SS ANOVA decomposition. Denote

$$\mathcal{H}^{(k)} = \mathcal{H}^{(k)}_{(0)} \oplus \mathcal{H}^{(k)}_{(1)},\ k = 1, 2,$$

where $\mathcal{H}^{(k)}_{(0)} = \{1\}$. Then in terms of the model space,

$$\begin{aligned} \mathcal{H} &\triangleq \mathcal{H}^{(1)} \otimes \mathcal{H}^{(2)} \\ &= \{\mathcal{H}^{(1)}_{(0)} \oplus \mathcal{H}^{(1)}_{(1)}\} \otimes \{\mathcal{H}^{(2)}_{(0)} \oplus \mathcal{H}^{(2)}_{(1)}\} \\ &= \{\mathcal{H}^{(1)}_{(0)} \otimes \mathcal{H}^{(2)}_{(0)}\} \oplus \{\mathcal{H}^{(1)}_{(0)} \otimes \mathcal{H}^{(2)}_{(1)}\} \oplus \{\mathcal{H}^{(1)}_{(1)} \otimes \mathcal{H}^{(2)}_{(0)}\} \oplus \{\mathcal{H}^{(1)}_{(1)} \otimes \mathcal{H}^{(2)}_{(1)}\} \\ &\triangleq \mathcal{H}_0 \oplus \mathcal{H}_x \oplus \mathcal{H}_y \oplus \mathcal{H}_{yx}, \end{aligned} \tag{2.13}$$

where $\mathcal{H}_0$, $\mathcal{H}_x$, $\mathcal{H}_y$ and $\mathcal{H}_{yx}$ are the functional spaces of μ, g_x, g_y and $g_{yx}(y, x)$ respectively.

A side condition is needed for the logistic transformation to be one-to-one. One possible side condition suggested by Gu (1995) is to set $\mu + g_x(x) = 0$ in (2.12). This side condition is equivalent to setting μ and g_x both equal to zero. Therefore an SS ANOVA model for g is

$$g = g_y(y) + g_{xy}(x, y), \tag{2.14}$$

and the model space $\mathcal{H}$ in (2.13) is reduced to

$$\mathcal{H}_{trim} \triangleq \mathcal{H}_y \oplus \mathcal{H}_{yx} \tag{2.15}$$

where $\mathcal{H}_y = \mathcal{H}_{(1)}^{(1)} \otimes \mathcal{H}_{(0)}^{(2)}$ and $\mathcal{H}_{yx} = \mathcal{H}_{(1)}^{(1)} \otimes \mathcal{H}_{(1)}^{(2)}$.

By applying a tensor sum decomposition, the functional space $\mathcal{H}_{trim}$ can be represented as

$$\mathcal{H}_{trim} = \mathcal{H}_{trim(0)} \oplus \mathcal{H}_{trim(1)}, \tag{2.16}$$

where $\mathcal{H}_{trim(0)}$ is a finite dimensional space with basis functions $\phi_1, ..., \phi_p$, and $\mathcal{H}_{trim(1)}$ is an RKHS with RK R_1.

Example: Tensor product cubic spline

Assume that $\mathcal{Y} = \mathcal{X} = [0, 1]$ and their corresponding marginal spaces $\mathcal{H}^{(1)} = \mathcal{H}^{(2)} = W_2^2$. The functional space of the joint function g is $W_2^2 \otimes W_2^2$. After applying SS ANOVA decomposition on $W_2^2 \otimes W_2^2$ and considering side conditions,

we have (2.15). Further decomposing $\mathcal{H}^{(k)}_{(1)} = \mathcal{H}^{(k)}_{(1,0)} \oplus \mathcal{H}^{(k)}_{(1,1)}$ for $k = 1, 2$, we have

$$\begin{aligned}
&\mathcal{H}_{trim(0)} \\
&= \{\mathcal{H}^{(1)}_{(1,0)} \otimes \mathcal{H}^{(2)}_{(0)}\} \oplus \{\mathcal{H}^{(1)}_{(1,0)} \otimes \mathcal{H}^{(2)}_{(1,0)}\}, \\
&\mathcal{H}_{trim(1)} \\
&= \{\mathcal{H}^{(1)}_{(1,1)} \otimes \mathcal{H}^{(2)}_{(0)}\} \oplus \{\mathcal{H}^{(1)}_{(1,1)} \otimes \mathcal{H}^{(2)}_{(1,0)}\} \oplus \{\mathcal{H}^{(1)}_{(1,0)} \otimes \mathcal{H}^{(2)}_{(1,1)}\} \oplus \{\mathcal{H}^{(1)}_{(1,1)} \otimes \mathcal{H}^{(2)}_{(1,1)}\},
\end{aligned}$$

where $\mathcal{H}^{(1)}_{(1,0)} = \{y - 0.5\}$, $\mathcal{H}^{(2)}_{(1,0)} = \{x - 0.5\}$, $\mathcal{H}^{(1)}_{(1,1)} = \{f \in W^2_2 : \int_0^1 f^{(i)}(y)dy = 0, i = 0, 1, \text{ and } \int_0^1 f^{(2)}(y)dy < \infty\}$, and $\mathcal{H}^{(2)}_{(1,1)} = \{f \in W^2_2 : \int_0^1 f^{(i)}(x)dx = 0, i = 0, 1, \text{ and } \int_0^1 f^{(2)}(x)dx < \infty\}$.

2.4.3 Penalized Likelihood Estimation

Consider finding the estimate of g in the trimmed space as defined in (2.16). Denote P_1 as the orthogonal projection operator onto $\mathcal{H}_{trim(1)}$. We find the estimate of g via minimizing the penalized likelihood

$$-\frac{1}{n}\sum_{i=1}^{n}\{g(Y_i, X_i) - \log\int_{\mathcal{Y}} e^{g(y,X_i)}dy\} + \frac{\lambda}{2}||P_1 g||^2, \tag{2.17}$$

in $\mathcal{H}_{trim}$.

The space $\mathcal{H}_{trim}$ might not be finite dimensional, hence the solution to (2.17) might not be computable. Denote $\mathbf{Z}_i = (Y_i, X_i)$ and set $\{\mathbf{Z}_{i_1}, ..., \mathbf{Z}_{i_q}\}$ as a random subset of $\mathbf{Z}_1, ..., \mathbf{Z}_n$. Gu (1995) proposed to find the solution in the following data adaptive finite dimensional space

$$\mathcal{H}_q = \mathcal{H}_{trim(0)} \oplus \{R_1(\mathbf{Z}_{i_j}, \cdot), j = 1, ..., q\}. \tag{2.18}$$

Define $\mu_g(h|x) = \int_{\mathcal{Y}} h(y,x)e^{g(y,x)}dy / \int_{\mathcal{Y}} e^{g(y,x)}dy$ and $V_g(h_1,h_2|x) = \mu_g(h_1h_2|x) - \mu_g(h_1|x)\mu_g(h_2|x)$. The solution to (2.17) in $\mathcal{H}_q$ can be obtained by the Newton updating equation which is similar to (2.6), with the $\mu_g(h)$ and $V_g(h_1,h_2)$ modified as follows,

$$\mu_g(h) = \frac{1}{n}\sum_{i=1}^{n}\mu_g(h|X_i), \qquad V_g(h_1,h_2) = \frac{1}{n}\sum_{i=1}^{n}V_g(h_1,h_2|X_i). \tag{2.19}$$

2.4.4 Smoothing Parameter Selection

Denote $f(x)$ as the marginal density of x. The aggregated relative K-L loss of $f_\lambda(y|x) = e^{g_\lambda}/\int_{\mathcal{Y}} e^{g_\lambda(y,x)}dy$ from $f(y|x) = e^{g}/\int_{\mathcal{Y}} e^{g(y,x)}dy$ is

$$RKL(g,g_\lambda) = \int_{\mathcal{X}} f(x)\left[\log\int_{\mathcal{Y}} e^{g_\lambda(y,x)}dy\right]dx - \int_{\mathcal{X}} f(x)\mu_g(g_\lambda|x)dx. \tag{2.20}$$

The first term of (2.20) can be estimated by $n^{-1}\sum_{i=1}^{n}\log\int_{\mathcal{Y}} e^{g_\lambda(y,X_i)}dy$. The second term of (2.20) can be estimated by the cross-validation sample mean $n^{-1}\sum_{i=1}^{n} g_\lambda^{[i]}(Y_i,X_i)$, where $g_\lambda^{[i]}(Y_i,X_i)$ minimizes a delete-one version of the quadratic approximation of (2.17) at $\widetilde{g} = g_\lambda$,

$$-\frac{1}{n-1}\sum_{j\neq i}^{n} g(Y_i,X_i) - \mu_{\widetilde{g}}(g) + \frac{1}{2}V_{\widetilde{g}}(g-\widetilde{g}, g-\widetilde{g}) + \frac{\lambda}{2}||P_1 g||^2,$$

for $\mu_g(h)$ and $V_g(h_1,h_2)$ in (2.19). The quadratic approximation of (2.17) at $\widetilde{g} = g_\lambda$ is obtained by applying the second-order Taylor expansion on $\log\int_{\mathcal{Y}} e^{g(y,X_i)}dy$.

The derivation of CV score for choosing the smoothing parameter λ follows the same procedure for the case of density estimation as mentioned in Section 2.3. For the detail about CV score for the case of conditional density, one may consult Gu (2013, Ch7).

Chapter 3

Nonparametric Mixed-Effects Density Regression for Repeated Measures Data

In a repeated measures design, data have a multilevel structure. Let ω represent a random subject in a population Ω with sampling distribution P_ω. The subject-specific joint probability density $f(y, x|\omega)$ is a random function on a product domain $\mathcal{Y} \times \mathcal{X} \times \Omega$. Let $f(y|x, \omega) = f(y, x|\omega)/\int_{\mathcal{Y}} f(y, x|\omega)dy$ be the subject-specific conditional density. Note that $f(y, x|\omega)$ and $f(y|x, \omega)$ are random since they both rely on a random sample ω. Now, assuming m subjects, $\omega_1, ..., \omega_m$, are sampled randomly from Ω. Let $(Y_{ij}, X_{ij}) \overset{iid}{\sim} f(y, x|\omega_i)$, $j = 1, ..., n_i$, be a sample from subject ω_i. Write $f(y|x, \omega_i) = f(y, x|\omega_i)/\int_{\mathcal{Y}} f(y, x|\omega_i)dy$ as the conditional density for the observed subject ω_i. The objective is to estimate $f(y|x, \omega_i)$ as well as to model the variation among all random subjects based on observations (Y_{ij}, X_{ij}). To model $f(y|x, \omega)$, we apply the logistic transformation,

$$f(y|x, \omega) = \frac{e^{g(y,x,\omega)}}{\int_{\mathcal{Y}} e^{g(y,x,\omega)}dy}. \tag{3.1}$$

We will consider model for $g(y, x,\omega)$. There are many ways to construct models for the multivariate random function $g(y, x,\omega)$, we will discuss one approach under the framework of Smoothing Spline ANOVA decompositions in the next section.

3.1 Smoothing Spline ANOVA (SS ANOVA) Decomposition

The SS ANOVA decomposition is an approach for building models for multivariate functions (Wahba, 1990; Gu and Wahba, 1991, 1993; Wahba *et al.*, 1995). It constructs functional spaces with hierarchical structure similar to the main effect and interactions in the classical ANOVA. Wang (2011) provides a concise introduction of SS ANOVA decomposition for various functional spaces. In the following subsection, five examples are provided to illustrate how to construct SS ANOVA decompositions for subject-specific logistic density $g(y, \omega)$ and subject-specific logistic conditional density $g(y, x, \omega)$.

3.1.1 Subject-Specific Density

For simplicity, set $\mathcal{Y} = [0, 1]$. Denote $g(y, \omega)$ as the subject-specific logistic density. Then the subject-specific density for subject ω is

$$f(y|\omega) = \frac{e^{g(y,\omega)}}{\int_{\mathcal{Y}} e^{g(y,\omega)} dy}. \tag{3.2}$$

With the sampled subject ω_i, the random samples $Y'_{ij}s$ are assumed i.i.d from $f(y|\omega_i)$. We will use linear and cubic splines to construct models for the subject-specific logistic density $g(y, \omega)$.

Linear Spline Density Model with Random Effects

Initially assume that the marginal function of y belongs to the Sobolev space W_2^1 for a linear spline. Define averaging operators A_1 and A_2 such that

$$\begin{aligned} A_1 g &= \int_\Omega g(y, \omega) dP_\omega, \\ A_2 g &= \int_0^1 g(y, \omega) dy. \end{aligned}$$

The SS ANOVA decomposition

$$\begin{aligned} g &= [A_1 + (I - A_1)][A_2 + (I - A_2)]g \\ &= A_1 A_2 g + A_1(I - A_2)g + (I - A_1)A_2 g + (I - A_1)(I - A_2)g \\ &\triangleq \mu + \gamma_1(y) + \phi(\omega) + \gamma_2(y, \omega)\,. \end{aligned} \tag{3.3}$$

The first two terms in (3.3) are fixed effects and orthogonal components in the RKHS W_2^1. The last two terms in (3.3) are random effects. The last term is an interaction between the subject ω and y.

Setting $\mu + \phi(\omega) = 0$ in (3.3) to ensure one-to-one logistic transform, we have an SS ANOVA model for g

$$g(y, \omega) = \gamma_1(y) + \gamma_2(y, \omega), \tag{3.4}$$

where the model space of $\gamma_1(y)$ is $W_2^1 \ominus \{1\}$ with RK $R(s,t) = k_1(s)k_1(t) + k_2(|s-t|)$, $\gamma_2(y,\omega)$ is a Gaussian process on $\mathcal{Y} \times \Omega$ with mean zero and covariance function $\sigma^2 R(s,t)$.

In (3.4), the functional fixed effect $\gamma_1(y)$ represents the overall mean logistic density and the functional random effect $\gamma_2(y,\omega)$ represents the subject-specific deviation.

Cubic Spline Density Model with Random Effects

Now assume the marginal function of y belongs to the Sobolev space W_2^2 for a cubic spline. Define averaging operators A_1 and A_2 and A_3 such that

$$\begin{aligned} A_1 g &= \int_\Omega g(y,\omega) dP_\omega, \\ A_2 g &= \int_0^1 g(y,\omega) dy, \\ A_3 g &= [\int_0^1 g'(y,\omega) dy](y - 0.5). \end{aligned}$$

The SS ANOVA decomposition

$$\begin{aligned} g &= [A_1 + (I - A_1)][A_2 + A_3 + (I - A_2 - A_3)]g \\ &= A_1 A_2 g + A_1 A_3 g + A_1(I - A_2 - A_3)g \\ &\quad + (I - A_1)A_2 g + (I - A_1)A_3 g + (I - A_1)(I - A_2 - A_3)g \\ &\triangleq \mu + \alpha \times (y - 0.5) + \gamma_1(y) + \phi(\omega) + \phi(\omega) \times (y - 0.5) + \gamma_2(y,\omega)\,. \quad (3.5) \end{aligned}$$

The first three terms in (3.5) are fixed effects and orthogonal components in the RKHS W_2^2. The last three terms in (3.5) are random effects. The last two terms are interactions between the subject ω and y.

Setting $\mu + \phi(\omega) = 0$ in (3.5) to ensure one-to-one logistic transform, we have an SS ANOVA model for g

$$g(y, \omega) = \alpha \times (y - 0.5) + \gamma_1(y) + \phi(\omega) \times (y - 0.5) + \gamma_2(y, \omega). \tag{3.6}$$

In (3.6), the model space of $\gamma_1(y)$ is $W_2^2 \ominus \{1, y-0.5\}$ with RK $R(s,t) = k_2(s)k_2(t) - k_4(|s-t|)$, $\phi(\omega)$ is drawn from $N(0, \sigma_1^2)$, $\gamma_2(y, \omega)$ is a Gaussian process on $\mathcal{Y} \times \Omega$ with mean zero and covariance function $\sigma_2^2 R(s,t)$. We assume that $\phi(\omega)$ and $\gamma_2(y, \omega)$ are independent of each other.

Define

$$g_1(y) = \alpha \times (y - 0.5) + \gamma_1(y),$$

and

$$g_2(y, \omega) = \phi(\omega) \times (y - 0.5) + \gamma_2(y, \omega).$$

Rewrite (3.6) as

$$g(y, \omega) = g_1(y) + g_2(y, \omega). \tag{3.7}$$

In (3.7), the functional fixed effect $g_1(y)$ represents the overall mean logistic density and the functional random effect $g_2(y, \omega)$ represents the subject-specific deviation.

3.1.2 Subject-Specific Conditional Density When Subjects Are Sampled From Multiple Populations

We use one example with linear and cubic spline models to illustrate model construction for the case when subjects are from different populations.

For simplicity, assume y is a continuous variable on $\mathcal{Y} = [0,1]$ and x is a discrete variable with domain $\mathcal{X} = \{1, ..., G\}$. Assume that ω is nested within level x. For every level x, denote Ω_x as the population from which subjects at level x are sampled with sampling distribution $P_{\omega|x}$. Under this scenario, the joint function $g(y, x, \omega)$ has domain $\mathcal{Y} \times \underset{x \in \mathcal{X}}{\cup} \{\{x\} \times \Omega_x\}$.We use linear and cubic spline to construct $g(y, x, \omega)$.

Linear Spline Conditional Density Model with Random Effects

Initially assume that the marginal function of y belongs to the Sobolev space W_2^1 for a linear spline. Also, let the model space for x be $\mathbb{R}^G$ where $\mathbb{R}^G$ is the Euclidean G-space. Define averaging operators A_1, A_2 and A_3 as follows:

$$\begin{aligned} A_1 g &= \int_0^1 g(y, x, \omega) dy, \\ A_3 g &= \int_{\Omega_x} g(y, x, \omega) dP_{\omega|x}, \\ A_2 g &= \frac{1}{G} \sum\nolimits_{x=1}^{G} A_3 g(y, x, \omega). \end{aligned}$$

An SS ANOVA decomposition can be defined as

$$\begin{aligned} g &= [A_1 + (I - A_1)][A_2 + (A_3 - A_2) + (I - A_3)]g \\ &= A_1 A_2 g + A_1(A_3 - A_2)g + A_1(I - A_3)g \\ &\quad + (I - A_1)A_2 g + (I - A_1)(A_3 - A_2)g + (I - A_1)(I - A_3)g \\ &\triangleq \mu + \beta(x) + \phi(x, \omega) + \gamma_1(y) + \gamma_2(y, x) + \gamma_3(y, x, \omega). \end{aligned} \tag{3.8}$$

To make the logistic transform one-to-one, we remove terms in (3.8) that do not dependent on y. Therefor, an SS ANOVA model for g is

$$g(y, x, \omega) = \gamma_1(y) + \gamma_2(y, x) + \gamma_3(y, x, \omega). \tag{3.9}$$

Denote two RKHS's

$$\begin{aligned} \mathcal{H}_y^{(1)} &= W_2^1 \ominus \{1\}, \\ \mathcal{H}_x^{(1)} &= R^G \ominus \{1\}, \end{aligned}$$

and their corresponding reproducing kernels (RKs)

$$\begin{aligned} R_1(s_1, t_1) &= k_1(s_1)k_1(t_1) + k_2(|s_1 - t_1|), \\ \tilde{R}_2(s_2, t_2) &= \delta_{s_2,t_2} - 1/G, \end{aligned}$$

where $\mathcal{H} \ominus \{1\}$ represents a RKHS $\mathcal{H}$ with constant functions been removed, $\otimes$ represents tensor product of RKHS's and $\delta_{u,v}$ is the Kronecker delta. Then the model space for γ_1 is $\mathcal{H}_y^{(1)}$ with RK R_1. The model space for γ_2 is $\mathcal{H}_y^{(1)} \otimes \mathcal{H}_x^{(1)}$ with RK $R_2((s_1, s_2), (t_1, t_2)) = R_1(s_1, t_1)\tilde{R}_2(s_2, t_2)$. Given a fixed level x, the random function $\gamma_3(y, x, \omega)$ is assumed to be a Gaussian process on $\mathcal{Y} \times \{\{x\}, \Omega_x\}$ with mean 0 and covariance function $\sigma_x^2 R_2((s, x), (t, x))$ where the parameter σ_x^2 depends on level x.

In (3.9), $\gamma_1(y)$ represents the average logistic density for all levels, $\gamma_2(y, x)$ represents departure of level x from the average logistic density $\gamma_1(y)$, and $\gamma_3(y, x, \omega)$ represents departure of subject ω from the level x logistic density $\gamma_1(y) + \gamma_2(y, x)$.

Cubic Spline Conditional Density Model with Random Effects

Let the marginal function of y belong to the Sobolev space W_2^2 for a cubic spline. In addition, let the model space for x be $\mathbb{R}^G$ where $\mathbb{R}^G$ is the Euclidean G-space. Define averaging operators A_1, A_2, A_3 and A_4 as follows:

$$\begin{aligned} A_1 g &= \int_0^1 g(y,x,\omega)dy, \\ A_2 g &= \{\int_0^1 g'(y,x,\omega)dy\}(y-0.5) \\ A_4 g &= \int_{\Omega_x} g(y,x,\omega)dP_{\omega|x}, \\ A_3 g &= \frac{1}{G}\sum\nolimits_{x=1}^{G} A_3 g(y,x,\omega). \end{aligned}$$

An SS ANOVA decomposition can be defined as

$$\begin{aligned} g &= [A_1 + A_2 + (I - A_1 - A_2)][A_3 + (A_4 - A_3) + (I - A_4)]g \\ &= A_1A_3g + A_1(A_4 - A_3)g + A_1(I - A_4)g \\ &\quad + A_2A_3g + A_2(A_4 - A_3)g + A_2(I - A_4)g \\ &\quad + (I - A_1 - A_2)A_3g + (I - A_1 - A_2)(A_4 - A_3)g + (I - A_1 - A_2)(I - A_4)g \\ &\triangleq \{\mu + \beta(x) + \phi(x,\omega)\} + (y - 0.5) \times \{\mu + \beta(x) + \phi(x,\omega)\} \\ &\quad + \gamma_1(y) + \gamma_2(y,x) + \gamma_3(y,x,\omega) \end{aligned} \tag{3.10}$$

To make the logistic transform one-to-one, we remove terms in (3.10) that do not dependent on$y$. Therefor, an SS ANOVA model for g is

$$g(y,x,\omega) = (y-0.5) \times \{\mu + \beta(x) + \phi(x,\omega)\} + \gamma_1(y) + \gamma_2(y,x) + \gamma_3(y,x,\omega). \tag{3.11}$$

Denote two RKHS's

$$\begin{aligned}\mathcal{H}_y^{(1)} &= W_2^2 \ominus \{1, y-0.5\},\\ \mathcal{H}_x^{(1)} &= R^G \ominus \{1\},\end{aligned}$$

and their corresponding reproducing kernels (RKs)

$$\begin{aligned}R_1(s_1,t_1) &= k_2(s_1)k_2(t_1) - k_4(|s_1-t_1|),\\ \tilde{R}_2(s_2,t_2) &= \delta_{s_2,t_2} - 1/G.\end{aligned}$$

Then the model space for $\gamma_1(y)$ is $\mathcal{H}_y^{(1)}$. The model space for $\gamma_2(y,x)$ is $\mathcal{H}_y^{(1)} \otimes \mathcal{H}_x^{(1)}$ with RK $R_2((s_1,s_2),(t_1,t_2)) = R_1(s_1,t_1)\tilde{R}_2(s_2,t_2)$. Given a fixed level x, the random function $\gamma_3(y,x,\omega)$ is assumed to be a Gaussian process on $\mathcal{Y}\times\{\{x\},\Omega_x\}$ with mean 0 and covariance function $\sigma_x^2 R_2((s,x),(t,x))$, where σ_x^2 depends on level x. In addition, the function $\beta(x)$ belongs to the functional space $\mathcal{H}_x^{(1)}$. For a fixed level x, we assume $\phi(x,\omega)$ is sampled from $N(0,\tau_x^2)$. Also, we assume $\gamma_3(y,x,\omega)$ and $\phi(x,\omega)$ are independent of each other.

Write

$$\begin{aligned}g_1(y) &= \mu\times(y-0.5) + \gamma_1(y),\\ g_2(y,x) &= \beta(x)\times(y-0.5) + \gamma_2(y,x),\end{aligned}$$

and

$$g_3(y,x,\omega) = \phi(x,\omega)\times(y-0.5) + \gamma_3(y,x,\omega).$$

Rewrite (3.11) as

$$g(y,x,\omega) = g_1(y) + g_2(y,x) + g_3(y,x,\omega). \tag{3.12}$$

Then $g_1(y)$ represents the average logistic density for all levels, $g_2(y,x)$ represents departure of level x from the average logistic density $g_1(y)$, and $g_3(y,x,\omega)$ represents departure of subject ω from the level x logistic density $g_1(y)+g_2(y,x)$.

3.1.3 Subject-Specific Conditional Density When Subjects Are Sampled From The Same Population

We only describe the case with linear spline model space for both y and x. A similar procedure can be performed for the cubic spline model space for both y and x.

Assume that y and x are both continuous variables with domains $\mathcal{Y}=[0,1]$ and $\mathcal{X}=[0,1]$. Denote Ω as the population from which the subjects $(\omega_i's)$ are sampled with sampling distribution P_ω. In this case, the joint function $g(y,x,\omega)$ has domain $\mathcal{Y}\times\mathcal{X}\times\Omega$. Suppose the marginal functions of y and x both belong to the Sobolev space W_2^1 for a linear spline. Define averaging operators A_1, A_2 and A_3 as follows:

$$\begin{aligned} A_1 g &= \int_0^1 g(y,x,\omega)dy, \\ A_2 g &= \int_\Omega g(y,x,\omega)dP_\omega, \\ A_3 g &= \int_0^1 g(y,x,\omega)dx. \end{aligned}$$

An SS ANOVA decomposition can be defined as

$$\begin{aligned} g &= [A_1 + (I - A_1)][A_2 + (I - A_2)][A_3 + (I - A_3)]g \\ &= A_1A_2A_3g + A_1A_2(I - A_3)g + A_1(I - A_2)A_3g \\ &\quad + A_1(I - A_2)(I - A_3)g + (I - A_1)A_2A_3g + (I - A_1)A_2(I - A_3)g \\ &\quad + (I - A_1)(I - A_2)A_3g + (I - A_1)(I - A_2)(I - A_3)g \\ &\triangleq \mu + \beta(x) + \phi_1(\omega) + \phi_2(x, \omega) + \gamma_1(y) + \gamma_2(y, x) + \gamma_3(y, \omega) + \gamma_4(y, x, \omega). \end{aligned}$$

We can make logistic transformation one-to-one by removing components that do not depend on y. Hence an SS ANOVA model for g is

$$g(y, x, \omega) = \gamma_1(y) + \gamma_2(y, x) + \gamma_3(y, \omega) + \gamma_4(y, x, \omega). \tag{3.13}$$

In (3.13), the model space for $\gamma_1(y)$ is $W_2^1 \ominus \{1\}$ and for $\gamma_2(y)$ is $(W_2^1 \ominus \{1\}) \otimes (W_2^1 \ominus \{1\})$ where $W_2^1 \ominus \{1\}$ and $(W_2^1 \ominus \{1\}) \otimes (W_2^1 \ominus \{1\})$ are RKHS's with RKs $R_1(s, t) = k_1(s)k_1(t)+k_2(|s-t|)$ and $R_2((s_1, s_2), (t_1, t_2)) = R_1(s_1, t_1)R_1(s_2, t_2)$. We assume that $\gamma_3(y, \omega)$ is a Gaussian process on $\mathcal{Y} \times \Omega$ with mean 0 and covariance function $\sigma_1^2 R_1(s, t)$. Similarly, with a fixed level x, we assume that $\gamma_4(y, x, \omega)$ is a Gaussian process on $\mathcal{Y}\times\Omega$ with mean 0 and covariance function $\sigma_2^2 R_2((s, x), (t, x))$.

In (3.13), $\gamma_1(y)$ represents the overall average logistic density, $\gamma_2(y, x)$ represents the interaction between y and x, $\gamma_3(y, \omega) + \gamma_4(y, x, \omega)$ represents departure of subject ω from $\gamma_1(y) + \gamma_2(y, x)$.

3.2 Nonparametric Mixed-Effects Density Regression

In the previous section we provided four examples to illustrate the application of SS ANOVA decomposition when building a model for the subject logistic conditional density $g(y, x, \omega)$. Other SS ANOVA models for general model space for y and general domain of x may be constructed similarly. Now we consider a more general model.

Assume that we have m subjects, $\omega_1, ..., \omega_m$, and that each subject ω_i generates a random sample of size n_i, $\{(Y_{ij}, X_{ij})\}_{j=1}^{n_i}$. Note that the selected subjects are allowed to be from different populations. Suppose the domains $\mathcal{Y}$ and $\mathcal{X}$ are arbitrary sets for generality. For subject ω_i, given a covariate $X_{ij} = x_{ij}$ and random effect $b_{ij} = \{b_i(y, x_{ij}) | y \in \mathcal{Y}\}$, Y_{ij} has the following density

$$f(y, x_{ij}, b_{ij}) = \frac{\exp\{\eta(y, x_{ij}, b_i(y, x_{ij}))\}}{\int_{\mathcal{Y}} \exp\{\eta(y, x_{ij}, b_i(y, x_{ij}))\} dy}.$$

Note that, rather than the joint density of y, x_{ij} and b_{ij}, $f(y, x_{ij}, b_{ij})$ represents density of Y_{ij} conditional on $X_{ij} = x_{ij}$ and b_{ij}. To model effects of covariates and variation among subjects, we propose the following nonparametric mixed-effects density regression (NMDR) model,

$$\eta(y, x_{ij}, b_i(y, x_{ij})) = \eta_1(y) + \eta_2(y, x_{ij}) + b_i(y, x_{ij}), \tag{3.14}$$

where η_1 and η_2 are fixed effects. We assume that η_1 and η_2 belong to RKHS's $\mathcal{H}_1$ with RK R_1 and $\mathcal{H}_2$ with RK R_2 respectively. In general, the random effects

$b'_{ij}s$ are assumed to be realizations of independent Gaussian processes with mean 0 and covariance function $\sigma(s,t|x_{ij})$.

The illustrated SS ANOVA models in the previous section are all special cases of (3.14). For example, in the case of the linear spline density model (3.4), we have $\eta_1(y) = \gamma_1(y)$, $\eta_2(y,x_{ij}) = 0$ and $b_i(y) = \gamma_2(y,\omega_i)$. In this case, the random effects merely depends on y and ω_i. Also, $\{b_i(y)|y \in \mathcal{Y}\}$, $i = 1,...,m$ are realizations of independent Gaussian processes with mean 0 and covariance function $\sigma(s,t) = \sigma^2 R(s,t)$. In the case of the cubic spline subject conditional density (3.12) where subjects are sampled from different populations, given $X_{ij} = x_{ij}$, we have $\eta_1(y) = g_1(y)$, $\eta_2(y,x_{ij}) = g_2(y,x_{ij})$ and $b_i(y,x_{ij}) = g_3(y,x_{ij},\omega_i)$. Furthermore, given a fixed level x_{ij}, $\{b_i(y,x_{ij})|y \in \mathcal{Y}\}$ is a realization of a Gaussian process with mean 0 and covariance function $\sigma(s,t|x_{ij}) = \tau^2_{x_{ij}} \times (s-0.5) \times (t-0.5) + \sigma^2_{x_{ij}} R_2((s,x_{ij}),(t,x_{ij}))$. In the case of the cubic spline subject conditional density (3.13) where subjects are sampled from the same population, given that $X_{ij} = x_{ij}$, we have $\eta_1(y) = \gamma_1(y)$, $\eta_2(y,x_{ij}) = \gamma_2(y,x_{ij})$ and $b_i(y,x_{ij}) = \gamma_3(y,\omega_i) + \gamma_4(y,x_{ij},\omega_i)$ where the collection $\{b_i(y,x_{ij})|y \in \mathcal{Y}\}$ is a realization of a Gaussian process with mean 0 and covariance function $\sigma(s,t|x_{ij}) = \sigma^2_{\mathbf{1}} R_1(s,t) + \sigma^2_{\mathbf{2}} R_2((s,x_{ij}),(t,x_{ij}))$.

In summary, for subject ω_i, conditional on $X_{ij} = x_{ij}$ and b_{ij}, Y_{ij} has the following mixed effects conditional density,

$$f(y,x_{ij},b_{ij}) = \frac{\exp\{\eta_1(y) + \eta_2(y,x_{ij}) + b_i(y,x_{ij})\}}{\int_{\mathcal{Y}} \exp\{\eta_1(y) + \eta_2(y,x_{ij}) + b_i(y,x_{ij})\} dy}. \tag{3.15}$$

The function (3.15) should be interpreted as the density of Y_{ij} conditional on the covariate x_{ij} and the random effects b_{ij}. The random effects introduce correlations within each subject. Given that $X_{ij} = x_{ij}$, the realization of a stochastic process $b_{ij} = \{b_i(y, x_{ij})|y \in \mathcal{Y}\}$ represents the interaction effect of covariate X and subject ω_i. To estimate nonparametric functions η_1 and η_2 and variance components associated with random effects, we utilize a penalized likelihood approach which will be introduced in the next chapter.

Chapter 4

Estimation and Computation for NMDR

4.1 Penalized Likelihood

Note, in a NMDR model (3.14), $\eta_1(y) \in \mathcal{H}_1$ with domain $\mathcal{Y}$ and $\eta_2(y,x) \in \mathcal{H}_2$ with domain $\mathcal{Y}\times\mathcal{X}$ where $\mathcal{H}_1$ and $\mathcal{H}_2$ are RKHS's. For $k = 1, 2$, we can decompose

$$\mathcal{H}_k = \mathcal{H}_k^0 \oplus \mathcal{H}_k^1, \tag{4.1}$$

where the subspace $\mathcal{H}_k^0 = span\{\phi_{kj},\ j = 1, ..., m_k\}$ is a finite dimensional space containing functions which are not penalized. The subspace $\mathcal{H}_k^1$ is the orthogonal complement of $\mathcal{H}_k^0$ in $\mathcal{H}_k$. Denote the reproducing kernel (RK) of $\mathcal{H}_k^1$ as $R_{k,1}$.

Denote $\mathbf{Y}_i = (Y_{i1,...,}Y_{in_i})^T$ as a vector that contains all Y_{ij}'s from subject ω_i. Given that $X_{ij} = x_{ij}$, let $B_{ij} = \{B_i(y, x_{ij}), y \in \mathcal{Y}\}$ be a stochastic process that generates the realization $b_{ij} = \{b_i(y, x_{ij}), y \in \mathcal{Y}\}$ and $\mathbf{B}_i$ as a collection of stochastic processes $\{B_{ij},\ j = 1, ...n_i\}$ that generates all random effects associated with subject ω_i. Let $p_{\mathbf{B}_i}$ and $p_{\mathbf{Y}_i|\mathbf{B}_i}$ be the probability density functions of $\mathbf{B}_i$ and $\mathbf{Y}_i$ conditional on $\mathbf{B}_i$ respectively. Depending on the domain $\mathcal{Y}$, one may use Radon–

Nikodym derivative or the finite-dimensional distributions property to construct the density function of $\mathbf{B}_i$. More detail about densities for stochastic processes can be found in Striebel (1959) and Barndorff-Nielsen and Sørensen (1994) .

We define the log marginal likelihood as

$$l(\zeta, \eta) = \sum_{i=1}^{m} \log E_{\mathbf{B}_i}[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)], \tag{4.2}$$

where $E_{\mathbf{B}_i}$ is with respect to the probability measure that governs the stochastic processes $\mathbf{B}_i$, ζ collects all parameters related to the random effects, $\eta = (\eta_1, \eta_2)$ collects all unknown nonparametric functions, and

$$p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) = \prod_{j=1}^{n_i} \frac{\exp\{\eta(Y_{ij}, X_{ij}, B_i(Y_{ij}, X_{ij}))\}}{\int_{\mathcal{Y}} \exp\{\eta(y, X_{ij}, B_i(y, X_{ij}))\} dy}. \tag{4.3}$$

Let the total sample size $N \triangleq \sum_{i=1}^{m} n_i$. Write $P_{k,1}$ as the projection operator onto the subspace $\mathcal{H}_k^1$ in $\mathcal{H}_k$ and $||P_{k,1}\eta_k||^2$ as a penalty on the departure from the null space $\mathcal{H}_k^0$. We estimate ζ and $\eta = (\eta_1, \eta_2)$ as minimizers of the penalized likelihood (PL)

$$PL = -\frac{1}{N} l(\zeta, \eta) + \frac{\lambda}{2} \sum_{k=1}^{2} \theta_k^{-1} ||P_{k,1}\eta_k||^2, \tag{4.4}$$

where λ and $\theta = (\theta_1, \theta_2)$ are smoothing parameters.

Writing g as the function representing the fixed effect in (3.14)

$$g(y, x) = \eta_1(y) + \eta_2(y, x). \tag{4.5}$$

Let $\mathcal{H} = \mathcal{H}^0 \oplus \mathcal{H}^1$ where $\mathcal{H}^0 = \mathcal{H}_1^0 \oplus \mathcal{H}_2^0$ and $\mathcal{H}^1 = \mathcal{H}_1^1 \oplus \mathcal{H}_2^1$. For any function $f \in \mathcal{H}^1$,

$$f(y, x) = f_1(y) + f_2(y, x),$$

where $f_k \in \mathcal{H}_k^1$, $k = 1, 2$. For any two functions f and $g \in \mathcal{H}^1$, define an inner product in $\mathcal{H}^1$ as

$$(f, g)_* = \sum_{k=1}^{2} \theta_k^{-1}(f_k, g_k), \tag{4.6}$$

where $f_k, g_k \in \mathcal{H}_k^1$, $k = 1, 2$. Hence, $||f||^2 = (f, f)_* = \sum_{k=1}^{2} \theta_k^{-1}||f_k||^2$. Let

$$R_\theta = \sum_{k=1}^{2} \theta_k R_{k,1}.$$

We have $R_\theta((y, x), (\cdot, \cdot)) \in \mathcal{H}^1$ and for any $f \in \mathcal{H}^1$,

$$\begin{aligned}
(R_\theta((y, x), (\cdot, \cdot)),\ f(\cdot, \cdot)) &= \theta_1^{-1}(\theta_1 R_{1,1}(y, \cdot),\ f_1(\cdot)) + \theta_2^{-1}(\theta_2 R_{2,1}((y, x), (\cdot, \cdot)),\ f_2(\cdot, \cdot)) \\
&= f_1(y) + f_2(y, x) \\
&= f(y, x),
\end{aligned}$$

thus R_θ is the RK of $\mathcal{H}^1$ with the inner product (4.6). Let $P_1^* = \sum_{k=1}^{2} P_{k,1}$ be the orthogonal projection in $\mathcal{H}$ onto $\mathcal{H}^1$. Then the penalized likelihood (4.4) is reduced to

$$PL = -\frac{1}{N} l(\zeta, g) + \frac{\lambda}{2}||P_1^* g||^2. \tag{4.7}$$

Now the goal is to estimate the vector of variance parameters ζ and fixed effect function g through minimizing (4.7).

4.2 Estimation for Fixed Effects

4.2.1 An Approximated Solution to the Penalized Likelihood

We now focus on finding the solution of g in (4.5) as the minimizer of the penalized likelihood in $\mathcal{H}$ when ζ is fixed. Usually, the space $\mathcal{H}$ is an infinite dimensional space. Hence the solution to the PL (4.7) in $\mathcal{H}$ is generally not computable.

Denote $\mathbf{Z}_{ij} = (Y_{ij}, X_{ij})^T$. We overcome the infinite dimensional problem by solving the minimization problem of (4.7) in the following data-adaptive space

$$\mathcal{H}^{(q)} = \mathcal{H}^0 \oplus span\{R_\theta(\mathbf{Z}_{(ij)_l}, \cdot); l = 1, ..., q\}, \tag{4.8}$$

where $\{\mathbf{Z}_{(ij)_1}, ..., \mathbf{Z}_{(ij)_q}\}$ is a random subset of observations $\mathbf{Z}_{ij}$, $j = 1, ..., n_i$, $i = 1, ..., m$. Gu and Wang (2003) suggested that a q closed to $10N^{2/9}$ is sufficient for a tensor product cubic spline logisitic density function without random effect in the sense that the estimate in the data-adaptive finite dimensional subspace $\mathcal{H}^{(q)}$ and $\mathcal{H}$ have the same convergence rate. For selecting $\mathbf{Z}_{(ij)_l}$, one may use simple random sampling for computational simplicity or stratified sampling for efficient estimation of variance components.

Denote $\phi_1, ..., \phi_p$ as basis functions of $\mathcal{H}^0$. The solution of g in $\mathcal{H}^{(q)}$ that minimizes (4.7) can be represented as

$$\hat{g}(\mathbf{z}) = \sum_{\nu=1}^{p} d_\nu \phi_\nu(\mathbf{z}) + \sum_{l=1}^{q} c_l R_\theta(\mathbf{Z}_{(ij)_l}, \mathbf{z}).$$

Denote $\mathbf{d} = (d_1, ..., d_p)^T$ and $\mathbf{c} = (c_1, ..., c_q)^T$. The solution in the vector form is

$$\hat{g} = \phi^T \mathbf{d} + \xi^T \mathbf{c}, \tag{4.9}$$

where ϕ =$(\phi_1, ..., \phi_p)^T$, $\xi = (\xi_1, ...,\xi_q)$ and $\xi_l = R_\theta(\mathbf{Z}_{(ij)_l}, \cdot)$.

Based on (4.9), the PL (4.7) can be rewritten as

$$PL(\zeta, \mathbf{c}, \mathbf{d}) = -\frac{1}{N} l(\zeta, \mathbf{c}, \mathbf{d}) + \frac{\lambda}{2} \mathbf{c}^T Q_\theta \mathbf{c}, \tag{4.10}$$

where Q_θ is a $q \times q$ matrix with the (k, l)th entry $R_\theta(\mathbf{Z}_{(ij)_k}, \mathbf{Z}_{(ij)_l})$.

4.2.2 Newton-Raphson Procedure

With ζ being fixed, coefficients $\mathbf{c}$ and $\mathbf{d}$ that minimize (4.10) are estimated from data through Newton-Raphson procedure.

Define

$$G_i \triangleq \frac{\partial \log p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)}{\partial (\mathbf{c}^T, \mathbf{d}^T)^T}$$

and

$$H_i \triangleq \frac{\partial^2 \log p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)}{\partial (\mathbf{c}^T, \mathbf{d}^T)^T \partial (\mathbf{c}^T, \mathbf{d}^T)}.$$

Taking the first two derivatives of the marginal likelihood $l(\zeta, \mathbf{c}, \mathbf{d})$, we have

$$\frac{\partial l(\zeta, \mathbf{c}, \mathbf{d})}{\partial (\mathbf{c}^T, \mathbf{d}^T)^T} = \sum_{i=1}^{m} E_{\mathbf{B}_i|\mathbf{Y}_i}(G_i), \tag{4.11}$$

$$\frac{\partial^2 l(\zeta, \mathbf{c}, \mathbf{d})}{\partial (\mathbf{c}^T, \mathbf{d}^T)^T \partial (\mathbf{c}^T, \mathbf{d}^T)} = \sum_{i=1}^{m} \{E_{\mathbf{B}_i|\mathbf{Y}_i}(H_i) + E_{\mathbf{B}_i|\mathbf{Y}_i}(G_i^2) - [E_{\mathbf{B}_i|\mathbf{Y}_i}(G_i)]^2\}. \tag{4.12}$$

The derivation of these derivatives can be found in Appendix A. The second order term $E_{\mathbf{B}_i|\mathbf{Y}_i}(G_i^2) - [E_{\mathbf{B}_i|\mathbf{Y}_i}(G_i)]^2$ in (4.12) can be dropped as suggested by

Benveniste, Metivier and Priouret (1987) and Jiang, Karcher and Wang (2011). As a result, one has

$$\frac{\partial^2 l(\zeta, \mathbf{c}, \mathbf{d})}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T \partial(\mathbf{c}^T, \mathbf{d}^T)} \approx \sum_{t=1}^{m} E_{\mathbf{B}_i|\mathbf{Y}_i}(H_i), \tag{4.13}$$

which is usually well behaved. For example, it is positive definite for convex target functions.

The first two derivatives of PL (4.10) are listed as follows,

$$\frac{\partial PL(\zeta, \mathbf{c}, \mathbf{d})}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} = -\frac{1}{N}\sum_{i=1}^{m} E_{\mathbf{B}_i|\mathbf{Y}_i}(G_i) + \frac{\lambda}{2}\frac{\partial \mathbf{c}^T Q_\theta \mathbf{c}}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T}, \tag{4.14}$$

and

$$\frac{\partial^2 PL(\zeta, \mathbf{c}, \mathbf{d})}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T \partial(\mathbf{c}^T, \mathbf{d}^T)} = -\frac{1}{N}\sum_{t=1}^{m} E_{\mathbf{B}_i|\mathbf{Y}_i}(H_i) + \frac{\lambda}{2}\frac{\partial^2 \mathbf{c}^T Q_\theta \mathbf{c}}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T \partial(\mathbf{c}^T, \mathbf{d}^T)}. \tag{4.15}$$

Let

$$\begin{aligned}
\mu_{h_1}(h_2|X_{ij}, B_{ij}) &\triangleq \int_{\mathcal{Y}} h_2(y, X_{ij}) \frac{e^{h_1(y,X_{ij})+B_i(y,X_{ij})}}{\int_{\mathcal{Y}} e^{h_1(y,X_{ij})+B_i(y,X_{ij})} dy} dy, \\
V_{h_1}(h_2, h_3|X_{ij}, B_{ij}) &\triangleq \mu_{h_1}(h_2 h_3|X_{ij}, B_{ij}) - \mu_{h1}(h_2|X_{ij}, B_{ij})\mu_{h2}(h_3|X_{ij}, B_{ij}), \\
V_{h_1}(h_2|X_{ij}, B_{ij}) &\triangleq V_{h_1}(h_2, h_2|X_{ij}, B_{ij}).
\end{aligned}$$

Set

$$\mu_{h_1}(h_2) = \frac{1}{N}\sum_{i=1}^{m} E_{\mathbf{B}_i|\mathbf{Y}_i}\{\sum_{j=1}^{n_i} \mu_{h_1}(h_2|X_{ij}, B_{ij})\}, \tag{4.16}$$

$$V_{h_1}(h_2, h_3) = \frac{1}{N}\sum_{i=1}^{m} E_{\mathbf{B}_i|\mathbf{Y}_i}\{\sum_{j=1}^{n_i} V_{h_1}(h_2, h_3|X_{ij}, B_{ij})\}. \tag{4.17}$$

Note, (4.16) and (4.17) require the expectation $E_{\mathbf{B}_i|\mathbf{Y}_i}\{\cdot\}$ for $i = 1, .., m$ which will be approximated by the MCMC method.

Denote U as the matrix that collects all $\mathbf{Z}'_{ij}s$, U= $(\mathbf{Z}_{11}, ..., \mathbf{Z}_{mn_m})^T$ with the k^{th} row denoted by $\mathbf{U}_k$. Also, let S be a $N \times p$ matrix with the $(j,k)th$ entry $\phi_k(\mathbf{U}_j)$ and R be an $N \times q$ matrix with the $(j,l)th$ entry $\xi_l(\mathbf{U}_j) = R_\theta(\mathbf{U}_{i_l}, \mathbf{U}_j)$ where $\mathbf{U}_{i_l}$ is a random subset of observations $\{\mathbf{U}_i, i = 1, ..., N\}$. Given the current estimate $\widetilde{g} = \phi^T\widetilde{\mathbf{d}} + \xi^T\widetilde{\mathbf{c}}$, the first and second derivatives (4.14) and (4.15) at $g = \widetilde{g}$ are

$$\begin{aligned}
\frac{\partial PL}{\partial \mathbf{d}} &= -\frac{1}{N}S^T\mathbf{1} + \mu_{\widetilde{g}}(\phi) = -\frac{1}{N}S^T\mathbf{1} + \mu_\phi, \\
\frac{\partial PL}{\partial \mathbf{c}} &= -\frac{1}{N}R^T\mathbf{1} + \mu_{\widetilde{g}}(\xi) + \lambda Q_\theta\widetilde{\mathbf{c}} = -\frac{1}{N}R^T\mathbf{1} + \mu_\xi + \lambda Q_\theta\widetilde{\mathbf{c}}, \\
\frac{\partial^2 PL}{\partial \mathbf{d}\partial \mathbf{d}^T} &= V_{\widetilde{g}}(\phi, \phi^T) = V_{\phi,\phi}, \\
\frac{\partial^2 PL}{\partial \mathbf{c}\partial \mathbf{c}^T} &= V_{\widetilde{g}}(\xi, \xi^T) = V_{\xi,\xi} + \lambda Q_\theta, \\
\frac{\partial^2 PL}{\partial \mathbf{d}\partial \mathbf{c}^T} &= V_{\widetilde{g}}(\phi, \xi^T) = V_{\phi,\xi}.
\end{aligned} \tag{4.18}$$

The Newton equation is thus

$$\begin{pmatrix} V_{\phi,\phi} & V_{\phi,\xi} \\ V_{\xi,\phi} & V_{\xi,\xi} + \lambda Q_\theta \end{pmatrix} \begin{pmatrix} \mathbf{d} \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} \frac{1}{N}S^T\mathbf{1} - \mu_\phi + V_{\phi,g} \\ \frac{1}{N}R^T\mathbf{1} - \mu_\xi + V_{\xi,g} \end{pmatrix}, \tag{4.19}$$

where $V_{\phi,g} = V_{\widetilde{g}}(\phi, \widetilde{g})$ and $V_{\xi,g} = V_{\widetilde{g}}(\xi, \widetilde{g})$.

4.3 Smoothing Parameter Selection

The smoothing parameters $\lambda\theta_1^{-1}$ and $\lambda\theta_2^{-1}$ are fixed in the Newton equation (4.19) during the updating procedure for estimating $\mathbf{c}$ and $\mathbf{d}$. In this section, we develop a data-driven approach to choose smoothing parameters. We evaluate the quality of an estimate f_λ with the Kullback-Leibler (K-L) loss. Since the K-L loss depends on unknown density f, we use cross-validation to estimate it.

4.3.1 Kullback-Leibler Loss

Let $\lambda = (\lambda\theta_1^{-1}, \lambda\theta_2^{-1})$. Write g_λ as the estimate of (4.5) obtained through minimizing penalized likelihood (4.10). For any subject ω, given a fixed covariate x and unobserved random effects $b_x = \{b(t,x)|t \in \mathcal{Y}\}$, we define the true and estimated subject conditional densities, $f(y|x,b_x)$ and $f_\lambda(y|x,b_x)$, as follows

$$\begin{aligned} f(y|x,b_x) &= \frac{\exp\{g(y,x)+b(y,x)\}}{\int_{\mathcal{Y}}\exp\{g(t,x)+b(t,x)\}dt}, \\ f_\lambda(y|x,b_x) &= \frac{\exp\{g_\lambda(y,x)+b(y,x)\}}{\int_{\mathcal{Y}}\exp\{g_\lambda(t,x)+b(t,x)\}dt}. \end{aligned}$$

Taking the expectation with respect to the measure that governs the stochastic process B_x that generates b_x, and weighting by the sampling proportion $f(x)$, the aggregated K-L loss of $f_\lambda(y|x,b_x)$ from $f(y|x,b_x)$ is

$$KL(f,f_\lambda) = \int_{\mathcal{X}} f(x)E_{B_x}\{\int_{\mathcal{Y}} f(y|x,B_x)[\log(\frac{f(y|x,B_x)}{f_\lambda(y|x,B_x)})dy]\}dx. \tag{4.20}$$

The relative K-L loss is

$$\begin{aligned} RKL(f,f_\lambda) &= \int_{\mathcal{X}} f(x)E_{B_x}\{\int_{\mathcal{Y}} f(y|x,B_x)[\log(\frac{1}{f_\lambda(y|x,B_x)})]dy\}dx \\ &= \int_{\mathcal{X}} f(x)E_{B_x}[\int_{\mathcal{Y}} f(y|x,B_x)\{\log\int_{\mathcal{Y}}\exp[g_\lambda(t,x)+B(t,x)]dt\}dy]dx \\ &\quad -\int_{\mathcal{X}} f(x)E_{B_x}[\int_{\mathcal{Y}} f(y|x,B_x)g_\lambda(y,x)dy]dx, \end{aligned} \tag{4.21}$$

where B_x is the stochastic process that generate the realization b_x.We select the smoothing parameter vector λ that minimize (4.21).

We now estimate the criterion (4.21) through data. The first term of (4.21) can be approximated by

$$\widehat{E}[\log \int_{\mathcal{Y}} e^{g_\lambda(t,X)+B(t,X)}dt] = \frac{1}{N}\sum\nolimits_{i=1}^{m}\sum\nolimits_{j=1}^{n_i} E_{B_{ij}}[\log \int_{\mathcal{Y}} e^{g_\lambda(t,X_{ij})+B_i(t,X_{ij})}dt], \tag{4.22}$$

where $E_{B_{ij}}$ is with respect to the measure that governs the stochastic process $B_{ij} = \{B_i(t, X_{ij})|t \in \mathcal{Y}\}$. The second term of (4.21) can be approximated by

$$\widehat{E}[g_\lambda(y,x)] = \frac{1}{N}\sum\nolimits_{i=1}^{m}\sum\nolimits_{j=1}^{n_i}[g_\lambda(Y_{ij},X_{ij})]. \tag{4.23}$$

However, using (4.23) usually leads to under-smoothing since we use the same data both for model fitting and validation. Standard cross-validation suggests to replace $g_\lambda(Y_{ij}, X_{ij})$ in (4.21) by $g_\lambda^{[(i,j)]}(Y_{ij}, X_{ij})$ which chosen to minimize the delete-one-observation version of (4.10).

In summary, the smoothing parameter selection criteria, (4.21) can be approximated by the following cross-validation estimate,

$$\begin{aligned} CV(\lambda) &= \frac{1}{N}\sum\nolimits_{i=1}^{m}\sum\nolimits_{j=1}^{n_i} E_{B_{ij}}[\log \int_{\mathcal{Y}} e^{g_\lambda(t,X_{ij})+B_i(t,X_{ij})}dt] \\ &\quad -\frac{1}{N}\sum\nolimits_{i=1}^{m}[\sum_{j=1}^{n_i} g_\lambda^{[(i,j)]}(Y_{ij},X_{ij})]. \end{aligned} \tag{4.24}$$

4.3.2 Cross-Validation

Computation of $g_\lambda^{[(i,j)]}(Y_{ij}, X_{ij})$ based on (4.24) for each $i = 1, ..., m$ and $j = 1, ..., n_i$ is costly, and we derive a more computationally efficient way to approximate it. The derivation follows the same steps as in Gu (2013 Ch7). Consider the following delete-one-observation version of a quadratic approximation to

(4.10) at $\widetilde{g}$,

$$-\frac{1}{N-1}\sum_{i=1}^{m}\sum_{k\neq j} g(Y_{ij},X_{ij}) + L_{\widetilde{g}} + \frac{\lambda}{2}\mathbf{c}^t Q_\theta \mathbf{c}, \tag{4.25}$$

where $L_{\widetilde{g}} = \mu_{\widetilde{g}}(g) - V_{\widetilde{g}}(\widetilde{g},g) + \frac{1}{2}V_{\widetilde{g}}(g,g)$ and $\mu_{\widetilde{g}}(\cdot)$ and $V_{\widetilde{g}}(\cdot,\cdot)$ are defined in Section 3.3.2. The derivative of quadratic approximation to the log marginal likelihood can be found in Appendix B. One should note that the terms $\mu_{\widetilde{g}}(\cdot)$ and $V_{\widetilde{g}}(\cdot,\cdot)$ in (4.25) use all data. In theory, we may use delete-one cluster. However, it would be very hard to derive a computable score if this approach were followed. Instead, we apply the delete-one observation only, so that a closed-form approximation to the true CV score can be obtained. Note that all we need is a good approximation to the true CV score.

Set $\widetilde{g} = g_\lambda$ in (4.25). Denote the resulting minimizer as $g_\lambda^{[i,j]}$. Let $\breve{\mathbf{c}} = (\mathbf{d}^T, \mathbf{c}^T)^T$ and $\breve{\xi} = (\phi^T, \xi^T)^T$. Rewrite (4.19) as $H\breve{\mathbf{c}} = \breve{R}^T\mathbf{1}/N + \mathbf{g}$, $\mathbf{g} = V_{\breve{\xi},\widetilde{\beta}} - \mu_{\breve{\xi}}$, H is the Hessian matrix appearing on the left side of (4.19), and $\breve{R}^T = (\breve{\xi}(Y_{11},X_{11}),...,\breve{\xi}(Y_{mn_m},X_{mn_m})) = (S,R)^T$. The minimizer $g_\lambda^{[i,j]}$ of (4.25) has the coefficient

$$\begin{aligned}\breve{\mathbf{c}}^{[i,j]} &= H^{-1}(\frac{\breve{R}^T\mathbf{1} - \breve{\xi}(Y_{ij},X_{ij})}{N-1} + \mathbf{g}) \\ &= \breve{\mathbf{c}} + \frac{1}{N(N-1)}H^{-1}\breve{R}^T\mathbf{1} - \frac{H^{-1}\breve{\xi}(Y_{ij},X_{ij})}{N-1}.\end{aligned} \tag{4.26}$$

Therefore

$$\begin{aligned}g_\lambda^{[i,j]}(Y_{ij},X_{ij}) &= \breve{\xi}(Y_{ij},X_{ij})^T\breve{\mathbf{c}}^{[i,j]} \\ &= \breve{\xi}(Y_{ij},X_{ij})^T\breve{\mathbf{c}} - \frac{1}{N-1}\breve{\xi}(Y_{ij},X_{ij})^T H^{-1}(\breve{\xi}(Y_{ij},X_{ij}) - \breve{R}^T\mathbf{1}/N).\end{aligned}$$

Noting that $\breve{R}^T = \sum_{i=1}^m \sum_{j=1}^{n_i} \breve{\xi}(Y_{ij}, X_{ij})$. The cross-validation estimator (4.23) can be computed by

$$\begin{aligned} \widehat{\mu_g}(g_\lambda) &= \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} g_\lambda^{[i,j]}(Y_{ij}, X_{ij}) \\ &= \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} g_\lambda(Y_{ij}, X_{ij}) - \frac{tr(P_1^\perp \breve{R}^T H^{-1} \breve{R}^T P_1^\perp)}{N(N-1)}, \end{aligned} \tag{4.27}$$

where $P_{\mathbf{1}}^\perp = I - \mathbf{1}\mathbf{1}^T/N$.

Plugging (4.27) into (4.24), we have the following approximate delete-one observation CV score,

$$\begin{aligned} CV_\alpha(\lambda) &= \frac{1}{N}\sum\nolimits_{i=1}^{m}\sum\nolimits_{j=1}^{n_i} E_{B_{ij}}[\log \int_{\mathcal{Y}} e^{g_\lambda(t,X_{ij})+B_i(t,X_{ij})}dt] \\ &\quad -\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} g_\lambda(Y_{ij}, X_{ij}) + \alpha\frac{tr(P_{\mathbf{1}}^\perp \breve{R}^T H^{-1} \breve{R}^T P_{\mathbf{1}}^\perp)}{N(N-1)}. \end{aligned} \tag{4.28}$$

The smoothing parameter therefore can be estimated as the minimizer of the CV score (4.28). The constant $\alpha > 1$ is added in (4.28) to prevent occasional undersmoothing. An α value around 1.4 was suggested for various density estimation problems; see Gu (2013, Ch7).

4.4 Estimation of Variance Component

In the NMDR model (3.15), given subjects $\omega_i' s$ and covariate $x_{ij}' s$, the unobserved random effects $b_{ij}' s$ are realizations of independent Gaussian processes with mean 0 and covariance function $\sigma(s, t|x_{ij})$. The covariance function $\sigma(s, t|x_{ij})$ can be modeled parametrically by assuming that it relies on a parsimonious set of parameters. It can also be modeled nonparametrically as discussed in Jennrich and

Schlucher (1986) for linear mixed-effect models and Rice and Silverman (1991) for data which are curves. In our research, we model the covariance structure parametrically by assuming that it relies on a parsimonious set of parameters ζ.

The vector ζ in (4.10) collects all parameters related the covariance structure of random effects. We estimate ζ through minimizing the PL (4.10) with $\mathbf{c}$ and $\mathbf{d}$ being fixed. Since the penalty term does not rely on ζ, we only need to minimize the negative profile likelihood, $-l(\zeta, \mathbf{c}, \mathbf{d})$. The first derivative is

$$\begin{aligned} -\frac{\partial}{\partial \zeta} l(\zeta, \mathbf{c}, \mathbf{d}) &= -\frac{\partial}{\partial \zeta} \sum_{i=1}^{m} \log E_{\mathbf{B}_i}[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)] \\ &= -\sum_{i=1}^{m} E_{\mathbf{B}_i|\mathbf{Y}_i}[\frac{\partial \log p_{\mathbf{B}_i}(\mathbf{B}_i;\zeta)}{\partial \zeta}]. \end{aligned} \tag{4.29}$$

And the second derivative is

$$\begin{aligned} -\frac{\partial^2}{\partial \zeta \partial \zeta^T} l(\zeta, \mathbf{c}, \mathbf{d}) &= -\sum_{i=1}^{m} \{E_{\mathbf{B}_i|\mathbf{Y}_i}[\frac{\partial^2 \log p_{\mathbf{B}_i}(\mathbf{B}_i;\zeta)}{\partial \zeta \partial \zeta^T}] + \\ & \qquad E_{\mathbf{B}_i|\mathbf{Y}_i}\{[\frac{\partial \log p_{\mathbf{B}_i}(\mathbf{B}_i;\zeta)}{\partial \zeta}]^2\} - \{E_{\mathbf{B}_i|\mathbf{Y}_i}[\frac{\partial \log p_{\mathbf{B}_i}(\mathbf{B}_i;\zeta)}{\partial \zeta}]\}^2\}, \end{aligned}$$

which can be approximated by

$$-\frac{\partial^2}{\partial \zeta \partial \zeta^T} l(\zeta, \mathbf{c}, \mathbf{d}) \approx -\sum_{i=1}^{m} E_{\mathbf{B}_i|\mathbf{Y}_i}[\frac{\partial^2 \log p_{\mathbf{B}_i}(\mathbf{B}_i;\zeta)}{\partial \zeta \partial \zeta^T}], \tag{4.30}$$

for the stability of computation as suggested by Benveniste, et al (1987) and Jiang, et al (2011). Thus, at the k^{th} iteration the updating equation is

$$\zeta^{(k)} = \zeta^{(k-1)} - [\sum\nolimits_{i=1}^{m} E^{(k-1)}_{\mathbf{B}_i|\mathbf{Y}_i}(D_{2,i}|_{\zeta=\zeta^{(k-1)}})]^{-1}[\sum\nolimits_{i=1}^{m} E^{(k-1)}_{\mathbf{B}_i|\mathbf{Y}_i}(D_{1,i}|_{\zeta=\zeta^{(k-1)}})], \tag{4.31}$$

where

$$\begin{aligned} D_{1,i} &= \frac{\partial \log p_{\mathbf{B}_i}(\mathbf{B}_i;\zeta)}{\partial \zeta}, \\ D_{2,i} &= \frac{\partial^2 \log p_{\mathbf{B}_i}(\mathbf{B}_i;\zeta)}{\partial \zeta \partial \zeta^T}, \end{aligned} \tag{4.32}$$

and $E^{(k-1)}_{\mathbf{B}_i|\mathbf{Y}_i}$ is the expectation that use the estimators obtained at the $(k-1)^{th}$ iteration.

4.5 Estimation Procedure

The estimation procedure contains the following sequence of steps. For fixed variance parameters ζ, we estimate fixed effects g in (4.5) by minimizing the PL in (4.7). The approximated solution in (4.9) can be calculated by the Newton-Raphson (N-R) procedure. The minimization is executed via two nested loops: for fixed smoothing parameter, the inner loop minimizes the PL in (4.10) through the N-R procedure; and the outer loop choose the optimal smoothing parameter by minimizing an approximation to the K-L loss based on a delete-one-observation CV score in (4.28). For the estimation of the variance parameter of the random effects, we find the MLE of ζ through minimizing PL in (4.10) with $\mathbf{c}$ and $\mathbf{d}$ being fixed.

The integrals with respect to the random effects involved in the Newton updating equations usually do not have closed forms. We approximate these integrals by using Markov Chain Monte Carlo (MCMC) sampling. In addition, the stochastic nature of MC sampling makes it difficult for the Newton updating procedure to converge to the optimum. We employ the Stochastic Approximation Algorithm (SAA) to control the sampling variation along iterations in the Newton updating procedure.

In the following, we first describe how to generate MCMC sample and then introduce SAA.

4.5.1 Markov Chain Monte Carlo

MCMC methods can be applied to generate samples when the target distribution is not easily sampled. We use Metropolis-Hastings (M-H) procedure to generate MCMC samples from the conditional distribution of $\mathbf{B}_i|\mathbf{Y}_i$. Additional information about M-H procedure can be found in Gelman et al. (2003) and Givens and Hoeting (2005). Our procedure is described as follows, where for notational convenience, we omit the i subscripts.

Denote $p(\mathbf{b}|\mathbf{y})$ as the conditional density of $\mathbf{B}|\mathbf{Y}$. We need a sample of size S from $p(\mathbf{b}|\mathbf{y})$. Given an initial value $\mathbf{b}^{(0)}$, we draw the sample using the following algorithm: for $l = 1, ..., S$,

1. Draw $\mathbf{b}^*$ from the proposal distribution $q(\mathbf{b}|\mathbf{b}^{(l-1)})$ and u from $U[0, 1]$;

2. Compute the M-H ratio r,

$$r = \frac{p(\mathbf{b}^*|\mathbf{Y})/q(\mathbf{b}^*|\mathbf{b}^{(l-1)})}{p(\mathbf{b}^{(l-1)}|\mathbf{Y})/q(\mathbf{b}^{(l-1)}|\mathbf{b}^*)}; \tag{4.33}$$

3. Set $\mathbf{b}^{(l)} = \mathbf{b}^*$ if $r > u$ and $\mathbf{b}^{(l)} = \mathbf{b}^{(l-1)}$ otherwise;

In our proposed model (3.15), the random effects $b'_{ij}s$ are realizations of independent Gaussian processes, hence we assume the distribution that generates the

collection $\mathbf{b}_i = \{b_{ij}, j = 1, ..., n_i\}$ is a multivariate normal distribution with density $p(\mathbf{b})$ if the stochastic process that associated with the random effect is finite dimensional. If the stochastic process that associated with the random effect is infinite dimensional, we can discretized the process and assume the collection of discretized processes also has a multivariate normal distribution. Assuming $p(\mathbf{b})$ is a multivariate normal density with mean $\mathbf{0}$ and covariance matrix Σ. A simple option for the proposal distribution, q, is to use a multivariate normal centered at the current sample, $\mathbf{b}^{(l-1)}$, with scaled covariance matrix $a^2\Sigma$. The constant a is chosen so that the acceptance rate is near 23% as suggested by Gelman et al. (2003, Ch11) for high dimensional MCMC sampling with Metropolis-Hastings procedure. Using multivariate normal proposal distribution simplifies the computation of the ratio r, since $q(\mathbf{b}^{(l-1)}|\mathbf{b}^*)$ and $q(\mathbf{b}^*|\mathbf{b}^{(l-1)})$ cancel in (4.33) and the ratio r is reduced to

$$\begin{aligned} r &= \frac{p(\mathbf{b}^*|\mathbf{Y})}{p(\mathbf{b}^{(l-1)}|\mathbf{Y})} \\ &= \frac{p(\mathbf{Y}|\mathbf{b}^*)p(\mathbf{b}^*)}{p(\mathbf{Y}|\mathbf{b}^{(l-1)})p(\mathbf{b}^{(l-1)})}. \end{aligned}$$

Strongly autocorrelated MCMC samples have a poor mixing property, which are unrepresentative of the true underlying target distribution. Christensen et al. (Ch.6 2011) suggest that MCMC samples with correlation for the observation that are 30 iterations apart as strongly autocorrelated. One may thin the strongly autocorrelated MCMC samples to have representative samples of the target dis-

tribution. For the multi-dimensional MCM chain in our case, we will check the univariate correlations for each dimension separately.

4.5.2 Stochastic Approximation Algorithm

The SAA was first proposed by Robbins and Monro (1951) for optimization problems where the objective function is given in a form of the expectation. Gu and Kong (1998, 2000) extended SAA for solving incomplete data estimation problems. See also Gu and Zhu (2001), Lai (2003) and Jiang, Karcher and Wang (2011).

Let $f_{\mathbf{e}}(\mathbf{e})$ be the density function of a random vector $\mathbf{e}$. Consider solving the following equation,

$$\mathbf{h}(\theta) = \mathbf{0}, \tag{4.34}$$

where θ is vector of parameters and $\mathbf{h}(\theta)$ is a vector valued function that can be written as the expectation of a function $\mathbf{H}(\theta, \mathbf{e})$, with respect to $\mathbf{e}$:

$$\mathbf{h}(\theta) = \int \mathbf{H}(\theta, \mathbf{e}) f_{\mathbf{e}}(\mathbf{e}) d\mathbf{e} = E_{\mathbf{e}}[\mathbf{H}(\theta, \mathbf{e})]. \tag{4.35}$$

In incomplete data estimation, $\mathbf{h}(\theta)$ usually is the first derivative with respect to θ of some criteria function such as marginal log-likelihood in generalized linear mixed-effects model (GLMM). The integrals with respect to $\mathbf{e}$ in (4.35) usually do not have closed analytic forms, hence solving equation (4.34) is very challenging. Monte Carlo (MC) sampling can be used to approximate the integral. But the new problem is that MC sampling's random nature leads to an algorithm that

may fail to converge to the optimum. One way to overcome this obstacle is using SAA which controls the sampling variation along iterations.

To apply the SAA, one needs to find a matrix $\mathbf{I}(\theta, \mathbf{e})$ such that $E_{\mathbf{e}}[\mathbf{I}(\theta, \mathbf{e})]$ is close to $\partial \mathbf{h}/\partial\theta$ in the neighborhood of the solution to (4.34). Benveniste, Metivier and Priouret (1987) proposed to use $\mathbf{I}(\theta, \mathbf{e}) = -\partial \mathbf{H}(\theta, \mathbf{e})/\partial\theta$ which is positive definite for convex target functions.

Let $\{\gamma_k, k \geq 1\}$ be a sequence of real numbers, and $\{m_k, k \geq 1\}$ be a sequence of positive integers which fulfill the following conditions:

C1. $\gamma_k \geq 1$ for all k,

C2. $\sum_{k=1}^{\infty} \gamma_k = 0$,

C3. $\sum_{k=1}^{\infty} \gamma_k^{1+\varepsilon}/m_k < \infty$ for some $\varepsilon \in (0, 1)$,

C4. $\sum_{k=1}^{\infty} |\gamma_k/m_k - \gamma_{k-1}/m_{k-1}| < \infty$.

At iteration k, an effective MCMC sample of size m_k, $\{\mathbf{e}^{(1)}, ..., \mathbf{e}^{(m_k)}\}$, with equilibrium distribution $f_{\mathbf{e}}(\mathbf{e})$ is drawn. The SAA updates the parameter vector θ and matrix $\mathbf{\Gamma}$ as follows:

$$\begin{aligned} \mathbf{\Gamma}_k &= (1 - \gamma_k)\mathbf{\Gamma}_{k-1} + \gamma_k \bar{\mathbf{I}}_k, \\ \theta_k &= \theta_{k-1} + \gamma_k \mathbf{\Gamma}_k^{-1} \overline{\mathbf{H}}_k, \end{aligned}$$

where

$$\overline{\mathbf{H}}_k = \frac{1}{m_k}\sum_{j=1}^{m_k}\mathbf{H}(\theta_{k-1}, \mathbf{e}_k^{(j)}),$$
$$\bar{\mathbf{I}}_k = \frac{1}{m_k}\sum_{j=1}^{m_k}\mathbf{I}(\theta_{k-1}, \mathbf{e}_k^{(j)}).$$

$\mathbf{\Gamma}_k$ behaves as an alternate of the Hessian matrix and is updated as a parameter matrix. γ_k is the step-size of the parameter updates. Convergence of the algorithm is guaranteed (Benveniste et al. 1987). In implementing SAA, m_k and γ_k need to satisfy conditions (C1)-(C4). Jiang, Karcher and Wang (2011) considered the following three combinations:

$$\text{G1. } \gamma_k = 1 \text{ and } m_k = m_0 + k^2,$$
$$\text{G2. } \gamma_k = 1/k \text{ and } m_k = m_0,$$
$$\text{G3. } \gamma_k = 1/\sqrt{k} \text{ and } m_k = m_0 + k,$$

where m_0 is the starting MCMC sample size.

4.5.3 Implemention

In our study, we apply the updating procedure (4.35) to solve equations (4.19) for $\mathbf{c}$ and $\mathbf{d}$ and (4.31) for variance parameter ζ.

In estimating $\mathbf{c}$ and $\mathbf{d}$, the goal is to solve

$$\frac{\partial PL(\zeta, \mathbf{c}, \mathbf{d})}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} = E_{\mathbf{B}|\mathbf{Y}}\{\mathbf{H}((\mathbf{c}, \mathbf{d}), \mathbf{B})\} = 0,$$

where

$$\mathbf{H}((\mathbf{c}, \mathbf{d}), \mathbf{B}) = \frac{-1}{N}\frac{\partial \log p_{\mathbf{Y}|\mathbf{B}}(\mathbf{Y})}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} + \frac{\partial}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T}\frac{\lambda}{2}\mathbf{c}^T Q\mathbf{c},$$

and the Hessian matrix is approximated by

$$I((\mathbf{c},\mathbf{d}),\mathbf{B}) = -\frac{\partial \mathbf{H}((\mathbf{c},\mathbf{d}),\mathbf{B})}{\partial(\mathbf{c}^T,\mathbf{d}^T)^T}.$$

In estimating ζ, we estimate ζ as the minimizers of the negative log marginal likelihood $-l(\zeta,\mathbf{c},\mathbf{d})$ in (4.2). The H matrix here is

$$H(\zeta,\mathbf{B}) = \frac{\partial \log p_{\mathbf{B}}(\mathbf{B};\zeta)}{\partial \zeta^T},$$

where $p_{\mathbf{B}}(\mathbf{B};\zeta)$ is multivariate normal density with mean $\mathbf{0}$ and variance parameters ζ, and the I matrix is

$$I(\zeta,\mathbf{B}) = \frac{\partial^2 \log p_{\mathbf{B}}(\mathbf{B};\zeta)}{\partial \zeta^T \partial \zeta}.$$

4.5.4 The Complete Algorithm

Gathering all pieces together, we have the following complete algorithm:

1. Provide initial values $\widehat{\mathbf{c}}^{(0)}, \widehat{\mathbf{d}}^{(0)}, \widehat{\zeta}^{(0)}$;

2. At iteration k

   (a) Draw a MCMC sample $\{\mathbf{b}_i^{(1)}, ..., \mathbf{b}_i^{(m_k)}\}$ for $i = 1, ..., m$ using the M-H procedure;

   (b) Updating $\mathbf{c}, \mathbf{d}$ by solving equation (4.19) with $E_{\mathbf{B}_i|\mathbf{Y}_i}[\mu_f(g|\mathbf{B}_i)]$ and $E_{\mathbf{B}_i|\mathbf{Y}_i}[V_f(g|\mathbf{B}_i)]$ computed by MCMC sample;

   (c) Draw another MCMC sample $\{\mathbf{b}_i^{(1)}, ..., \mathbf{b}_i^{(m_k)}\}$ for $i = 1, ..., m$ with updated $\mathbf{c}, \mathbf{d}$ and update ζ by equation (4.31) using SAA;

3. Repeat Step 2 until convergence.

Initial Value and Stopping Criterion

The initial values $(\widehat{\mathbf{c}}^{(0)}, \widehat{\mathbf{d}}^{(0)})$ and $\widehat{\zeta}^{(0)}$ can be any reasonable user-specified values. Denote $(\widehat{\mathbf{c}}^{*}, \widehat{\mathbf{d}}^{*})$ as the estimator found based on pooling all data together and ignoring subject effects in estimation. One may use $(\widehat{\mathbf{c}}^{*}, \widehat{\mathbf{d}}^{*})$ for initial values $(\widehat{\mathbf{c}}^{(0)}, \widehat{\mathbf{d}}^{(0)})$. For the initial value $\widehat{\zeta}^{(0)}$, we propose an initial value, $\widehat{\zeta}_{GM}$, with its computation described as follows. Denote $\widehat{g}_i$ as the individual logistic density estimator which merely uses data from subject ω_i in estimation and g as the true fixed effects for NMDR model. Let $\widehat{\mathbf{g}}_i$ and $\widehat{\mathbf{g}}$ represent the vectors of functions $\hat{g}_i$ and g evaluated at the grid points respectively. According to the settings of our proposed model (3.14), we assume $\widehat{\mathbf{g}}_i's$ are multivariate normal distributed with mean $\mathbf{g}$ and covariance matrix Σ_ζ. We can compute the estimater of ζ using MLE method with Gaussian process realizations $\widehat{\mathbf{g}}$. We call this estimate as GM estimate and denote it as $\widehat{\zeta}_{GM}$, since it is obtained by using the MLE of the mean function in Gaussian process.

To incorporate SAA in the algorithm, one need to update the proxy of the Hessian matrix $\mathbf{\Gamma}$, a simple choice for $\mathbf{\Gamma}^{(0)}$ is the identity matrix.

The convergence of the estimation of ($\mathbf{c}$, $\mathbf{d}$) is usually fast, but it usually takes longer for ζ to converge. Denote ε as the predetermined accuracy tolerance. Define the relative difference of estimates of ζ at the k^{th} iteration as,

$$d_\zeta^{(k)} = \frac{||\zeta^{(k)} - \zeta^{(k-1)}||}{||\zeta^{(k-1)}||},$$

where $||\cdot||$ is Euclidean distance. The loop stops at the k^{th} iteration if $d_{\zeta}^{(k)} < \epsilon$, where ϵ is user-specified. Other stopping rule can be found in Booth and Hobert (1999).

Chapter 5

Simulations

In this chapter, we conduct extensive simulations to evaluate the performance of the proposed methods. We generate data from known models and use our proposed NMDR model to estimate subject densities and variation among them. The simulation results indicates our methods perform well for the estimation both of densities and other variation among them.

We use the model introduced in Chapter 3 to generate data and then use the estimation methods developed in Chapter 4 to estimate the fixed effects and variance parameters. We assess the estimate of population density by using K-L loss, and we evaluate the variance parameter estimation performance via by using mean squared error (MSE).

5.1 Simulation Methods

5.1.1 Model for Generating Data

We use model (3.4) to generate data. Denote Y_{ij} as the the j^{th} observation from subject ω_i, $j = 1, ..., n_i$, $i = 1, ..., m$. The sample $\{Y_{ij}\}_{j=1}^{n_i}$ from subject ω_i is simulated from the following conditional density (in the form of NMDR model),

$$f(y, b_i) = \frac{\exp\{\eta_1(y) + b_i(y)\}}{\int_{\mathcal{Y}} \exp\{\eta_1(y) + b_i(y)\}dy}, \tag{5.1}$$

where $\eta_1(y) = -\frac{(y-\theta)^2}{2\tau}$ (for chosen θ and τ values described in the next paragraph) and where $b_i = \{b_i(y)|y \in \mathcal{Y}\}$ is a realization of a Gaussian process with mean 0 and covariance function $\sigma^2 R_1$ where $R_1(s,t) = k_1(s)k_1(t) - k_2(|s-t|)$. For simplicity, the domain $\mathcal{Y}$ is assumed to be interval $[0, 1]$.

When generating data, we discretize the domain $\mathcal{Y}$ into 200 equal length subdivisions, $I_1, ..., I_{200}$. Denote the center point of the kth subdivision as u_k, and $\mathcal{Y}^\star$=$\{u_1, ..., u_{200}\}$. For subject ω_i, we draw random samples from $\mathcal{Y}^\star$ with replacement and the probability of u_k being selected as π_{ik}=$\int_{I_k} f(y, b_i)dy$. The data drawn from $\mathcal{Y}^\star$ are viewed as the raw data from the true domain $\mathcal{Y}$.

On the domain $\mathcal{Y} = [0, 1]$, the shape of density function can be determined by the values of θ and τ. Different values of θ are set for skewed and symmetric population densities. The Gaussian process is infinite dimensional, so we use multivariate normal to generate realizations of the process, b_i, based on the discretized domain. Figure 5.1 shows 30 subject-specific densities and their population densities for a symmetric case with $\theta = 1/2$ and a skewed case with $\theta = 1/4$, for two

different values of $\sigma^2 = 0.5$ and 2. The value of τ is set to be $1/6$ throughout this chapter. The population density is defined as the density without the random effects,

$$f(y) = \frac{\exp\{\eta_1(y)\}}{\int_{\mathcal{Y}} \exp\{\eta_1(y)\} dy}.$$

As expected the variation among subject densities is larger when σ^2 is larger.

In the simulation study, the sample size for each subject n_i is set to be 200 for all experiments. One should note that n_i and number of subdivisions for data generating do not need to be equal, although the same number is used in our simulation (both are 200). For each subject, after 200 observations were simulated, we then bin the simulated data. Denote q as the number of bins. We consider a factorial design with the following choices of simulation parameters: $(\theta, \tau) = (1/4, 1/6)$ and $(1/2, 1/6)$, $m = 13$, 30 and 100, $\sigma^2 = 0.5$ and 2 and $q = 5$ and 10. In total, there are 24 combinations.

5.1.2 Estimation

Model

We fit the simulated data using NMDR with linear spline to model the subject densities and the variation among them. We use linear spline representers evaluated at middle points of bins. Specifically, the main effect is approximated by $\eta_1(y) = \sum_{l=1}^{q} c_l R_1(K_l, y)$, where K_l is the middle point of the l^{th} bin and the

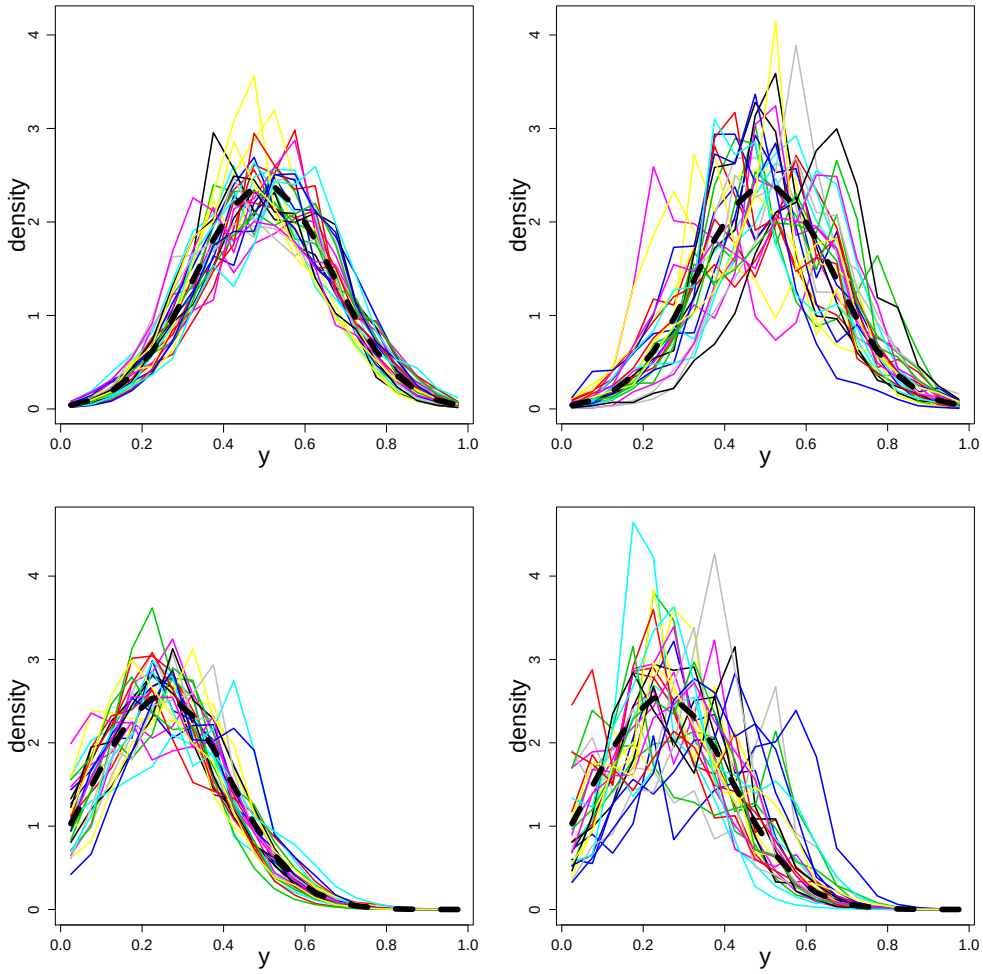


Figure 5.1: Subject-specific simulated densities: black line represents the population density (density without random effects), colored lines represent the subject-specific densities. The first row displays symmetric cases ($\theta = 1/2$) and the second row displays the skewed cases ($\theta = 1/4$). The left column corresponds to σ^2=0.5 and the right column corresponds to σ^2=2.

variation among subjects are modeled using zero-mean Gaussian processes with covariance function $\sigma^2 R_1$ where $R_1(s,t) = k_1(s)k_1(t) - k_2(|s-t|)$.

Updating Equation

We use Newton equation (4.19) to update $\hat{\eta}_1$. The expectation $E_{\mathbf{B}_i|\mathbf{Y}_i}(\cdot)$ in (4.19) is approximated using MCMC samples generated from Metropolis-Hastings algorithm. We use the updating equation (4.31) for updating $\hat{\sigma}^2$. Given the same stopping rule, the updating procedure of σ^2 usually needs much more iterations and hence more time than the updating procedure of η_1 to get convergence. One way to get fast convergence of the updating process for σ^2 is to use MCMC effective samples with large sample size to approximate integrals in updating equation at each iteration. The effective MCMC samples represents the samples that are obtained by discarded part of original MCMC samples for the goal of good mixing. The qualities that are described by the mixing properties of the MCMC sample can be found in Givens and Hoeting (2005).

Denote $h(\mathbf{B}_i)$ as any function of random effects $\mathbf{B}_i$. To make the updating procedure more efficient for $\hat{\sigma}^2$, we transform the conditional expectation $E_{\mathbf{B}_i|\mathbf{Y}_i}[h(\mathbf{B}_i)]$ to the expectation $E_{\mathbf{B}_i}(w_i h(\mathbf{B}_i))$ where the weight,

$$w_i = p(\mathbf{Y}_i|\mathbf{B}_i) / \int p(\mathbf{Y}_i|\mathbf{b}_i)p(\mathbf{b}_i;\sigma^2)d\mathbf{b}_i.$$

Hence the updating equation (4.31) is modified as

$$\hat{\sigma}^{2(k)}=\hat{\sigma}^{2(k-1)} - [\sum_{i=1}^{m} E_{\mathbf{B}_i}^{(k-1)}(w_i D_{\,2,i}|_{\sigma^2=\hat{\sigma}^{2(k-1)}})]^{-1} \sum_{i=1}^{m} E_{\mathbf{B}_i}^{(k-1)}(w_i D_{\,1,i}|_{\sigma^2=\hat{\sigma}^{2(k-1)}}), \tag{5.2}$$

where $D_{\,1,i}$ and $D_{\,2,i}$ are the first two derivatives of log likelihood function of variance component for the i^{th} subject in (4.32) with ζ being replaced by σ^2. The expectations $E_{\mathbf{B}_i}^{(k-1)}$ in (5.2) are approximated by Monte Carlo (MC) samples drawn directly from the distribution of random effects $p(\mathbf{b}_i; \hat{\sigma}^{2(k-1)})$. Given the fact that $p(\mathbf{b}_i; \hat{\sigma}^{2(k-1)})$ is a multivariate normal density, the MC samples can be generated easily.

Implementation in R

For the subject-specific density model introduced in Section 3.1.1, one may leverage the *R* function *ssden* in the library **gss** developed by Gu (2009) to estimate **c** and **d**. The *R* function *ssden* is developed for density estimation using SS ANOVA models. It includes an argument "bias" for input for sampling bias. Sampling bias is a bias in which the observations were collected in a biased way. One can save the exponential of MCMC samples, $e^{b_i(y)}$, to the parameter "bias" in function *ssden* to obtain a NMDR estimator for fixed-effect function in each iteration during the updating procedure. Applying *ssden* on the subject-specific density model only results in NMDR estimates with scheme G1 in (4.36). For other schemes one may need to modify *ssden*.

We now describe briefly the reason that one can leverage *ssden* for biased sampling to estimate NMDR model. The main point is that the Newton equation for estimating **c** and **d** for biased sampling and the one for NMDR model are the same. Consider the density for biased sampling as $f(y) = w(t,y)e^{g(y)}[\int_{\mathcal{Y}} w(t,y)e^{g(y)}dy]^{-1}$ where the weight function $w(t,y) \geq 0$ depends on t, the index of source that the observation y come from. Additional information about the index t can be seen in Gu (2013, Ch7). In leveraging the *ssden* for our case, we can treat the exponential of random effects, $e^{b_i(y)}$, as the weight function $w(t,y)$. The Newton equation for biased sampling is similar to the one for typical density estimation (2.6) but with modified $\mu_g(h)$ and $V_g(h_1,h_2)$ as

$$\begin{aligned} \mu_g(h) &= \frac{1}{n}\sum\nolimits_{i=1}^{n} \mu_g(h|t_i), \\ V_g(h,h') &= \frac{1}{n}\sum\nolimits_{i=1}^{n} v_g(h_1,h_2|t_i), \end{aligned}$$

where

$$\begin{aligned} \mu_g(h|t) &= \int_{\mathcal{Y}} h(y)w(t,y)\exp\{g(y)\}dy \Big/ \int_{\mathcal{Y}} w(t,y)\exp\{g(y)\}dy, \\ v_g(h_1,h_2|t) &= \mu_g(h_1h_2|t) - \mu_g(h_1|t)\mu_g(h_2|t). \end{aligned}$$

With modified $\mu_g(h)$ and $V_g(h_1,h_2)$, the Newton equation for biased sampling now is in the same form as (4.19) with X_i being removed from the equation. One should note that the biased sampling density estimation uses penalized likelihood for model estimation criteria, while NMDR model use penalized marginal likelihood. This fact does not affect the same form of the Newton equations for each of the biased sampling density estimation and NMDR model. But this

fact does affect the computation of CV score. For the subject-specific density model, the second term in the delete-one observation CV score (4.28) is simplified to $N^{-1}\sum_{=1}^{m} n_i E_{\mathbf{B_i}}[\log \int_{\mathcal{Y}} e^{g_\lambda(t)+B_i(t)}dt]$. However, if we leverage *ssden* to estimate the fixed effect, the expectation $E_{\mathbf{B_i}}[\log \int_{\mathcal{Y}} e^{g_\lambda(t)+B_i(t)}dt]$ in CV score is actually computed by $E_{\mathbf{B_i}|\mathbf{Y}_i}[\log \int_{\mathcal{Y}} e^{g_\lambda(t)+B_i(t)}dt]$. Note that $E_{\mathbf{B_i}|\mathbf{Y}_i}$ is the expectation of the random effects (a Gaussian process) conditional on the observations $\mathbf{Y}_i = (Y_{i1}, ..., Y_{in_i})$.

Computation setting

The maximum number of iterations for the whole updating procedure is set to be 75. The SAA algorithm with G1 setting which allows the effective MCMC and MC sample size increases quadratically along iterations is employed within the updating procedure. We start SAA algorithm from the 38^{th} iteration for updating $\hat{\eta}_1$ and at the beginning for updating $\hat{\sigma}^2$. Set S_1 and S_2 as initial effective MCMC sample size mad MC sample size for updating $\hat{\eta}_1$ and $\hat{\sigma}^2$ respectively. In other words, at the kth iteration the effective MCMC sample size for estimating η_1 is $S_1 + [(k-37)I_{\{k>37\}}]^2$ where I is an indicator function here, and the MC sample size for estimating σ^2 is $S_2 + k^2$. With σ^2 being fixed, the estimation of the main effect η_1 can be executed by using R function *ssden*.

Initial Value and Stopping Rule

Denote $\hat{\eta}_{1pooled}$ as the pooled estimates of η_1 which is an estimate that based on all observations and ignores subject effects. For the main effect, we set the initial values $\hat{\eta}_1^{(0)} = \hat{\eta}_{1pooled}$. For the variation among subjects, we set $\hat{\sigma}^{2(0)} = 1.5\sigma^2$. To measure the convergence of $\hat{\eta}_1$, we use the relative difference of estimates evaluated at the middle point of each bin $K_1, ..., K_q$. Define the relative difference of estimates of η_1 and σ^2 as follows,

$$\begin{aligned} d_{\eta_1} &= \frac{\sqrt{\sum_{l=1}^{q}[\hat{\eta}_1^{(i)}(K_l) - \hat{\eta}_1^{(i-1)}(K_l)]^2}}{\sqrt{\sum_{l=1}^{q}[\hat{\eta}_1^{(i-1)}(K_l)]^2}}, \\ d_{\sigma^2} &= \frac{|\hat{\sigma}^{2(i)} - \hat{\sigma}^{2(i-1)}|}{|\hat{\sigma}^{2(i-1)}|}. \end{aligned}$$

In each simulation, we let the updating procedure run at least 25 iterations. After the 25^{th} iteration, we set $\hat{\eta}_1^{(i)} = \hat{\eta}_1^{(i-1)}$ if $d_{\eta_1} < 10^{-5}$ and stop the procedure when $d_{\sigma^2} < 5 \times 10^{-4}$.

Output

We run $S = 100$ simulations for each experiment. We present the mean of $\hat{\sigma}^2$, the mean of K-L loss $KL(f, \hat{f})$ and MSE of $\hat{\sigma}^2$ to summarize our results. The mean of $\hat{\sigma}^2$ is computed by $\sum_{i=1}^{S} \hat{\sigma}_i^2/S$ where $\hat{\sigma}_i^2$ is the estimate from the result of the i^{th} simulation.The K-L loss between true population density $f(y) = e^{-\frac{1}{2}(\frac{y-\theta}{\tau})^2} / \int_0^1 e^{-\frac{1}{2}(\frac{y-\theta}{\tau})^2} dy$ and its estimate $\hat{f}(y) = e^{\hat{\eta}_1(y)} / \int_0^1 e^{\hat{\eta}_1(y)} dy$ is defined as

follows

$$KL(f, \hat{f}) = \int_0^1 \log \frac{f(y)}{\hat{f}(y)} f(y) dy,$$

which will be approximated by $\sum_{l=1}^{q} \log[f(K_l)/\hat{f}(K_l)] f(K_l)/q$. We will use the mean of K-L loss across 100 simulations to measure the accuracy of fitting population density. Finally, the MSE of $\hat{\sigma}^2$ is computed by

$$MSE(\hat{\sigma}^2) = \frac{\sum_{i=1}^{S} (\hat{\sigma}_i^2 - \sigma_i^2)^2}{S}.$$

5.1.3 MCMC sample

Proposal Distribution in Metropolis-Hastings algorithm

In generating MCMC samples for updating $\hat{\eta}_1$, at the i^{th} Newton updating iteration and the t^{th} MCMC iteration, the proposal distribution is multivariate normal with mean $\mathbf{b}^{(i,t-1)}$ and covariance matrix $a^2\hat{\Sigma}^{(i)}$ where $\mathbf{b}^{(i,t-1)}$ is generated at the $(t-1)^{th}$ MCMC iteration and $\hat{\Sigma}^{(i)} = \hat{\sigma}^{2(i)} R_1$.

A good value of a provides high quality MCMC samples which lead to more accurate approximation. The value a depends on the shape of the target distribution which is influenced by θ, τ and σ^2 and the number of subintervals q. We fix $\tau = 1/6$ and search for optimal a for each of our different combinations of θ, σ^2 and q. For each combination of θ, σ^2 and q with different a we repeat 1000 experiments. We choose a value of a that has median acceptance rate over 1000 experiments around 23%. In each experiment, we draw MCMC samples with sample size $10,000$ based on simulated observations and record the acceptance rate

at the end of each experiment. The MCMC samples here refers to the samples that before discarding part of samples for the goal of good mixing. Table 5.1 and 5.2 are example searching results which provide a values that will be used in the following simulations. In our simulation setting, the impact of σ^2 on a search results was found to be small, we thus select our a value only based on θ, τ and q throughout our simulations.

q	θ	a	Minimum	1st Quartile	Median	3rd Quartile	Maximum
10	0.25	0.52	20.06%	22.46%	23.22%	23.98%	26.62%
10	0.5	0.50	19.40%	22.45%	23.28%	24.12%	29.09%
5	0.25	0.38	18.11%	21.92%	22.89%	24.01%	28.79%
5	0.5	0.36	18.31%	21.52%	22.45%	23.46%	29.13%

Table 5.1: Searching result for a when $\tau = 1/6$ and $\sigma^2 = 2$.

q	θ	a	Minimum	1st Quartile	Median	3rd Quartile	Maximum
10	0.25	0.52	21.07%	22.53%	23.06%	23.60%	25.52%
10	0.5	0.50	20.53%	22.61%	23.17%	23.71%	26.06%
5	0.25	0.38	19.69%	22.10%	22.75%	23.44%	26.65%
5	0.5	0.36	18.97%	21.63%	22.29%	23.97%	25.32%

Table 5.2: Searching result for a when $\tau = 1/6$ and $\sigma^2 = 0.5$.

MCMC setting for Newton updating equation

For updating $\hat{\eta}_1$, the effective MCMC samples are obtained by storing every 10^{th} MCMC sample after an initial burn-in of 200 sweeps. We provide some results based on $1,000$ effective MCMC samples to show the way we choose effective MCMC samples and the Metropolis-Hastings proposal distribution we proposed in the previous section do produce effective MCMC samples with good mixing

(based on trace plots) and low autocorrelation (based on ACF plots). Figures 5.2-5.4 display the diagnostic plots for $\sigma^2 = 2$, 3 and 1 for the case that the Gaussian process is discretized as a 5 dimensional multivariate normal random vector $(B_1, ..., B_5)$ for the skewed density case $\theta = 1/4$. Figure 5.2 displays univariate trace and ACF plots for $\sigma^2 = 2$ which is one of the true values in our simulation settings. Figure 5.3 displays univariate trace and ACF plots for $\sigma^2 = 3$ which is used as the initial value when the true value $\sigma^2 = 2$. Since our simulation result usually indicates the convergence value is 0.5 times the true one, we provide Figure 5.4 for $\sigma^2 = 1$ which is the rough convergence value for the true value $\sigma^2 = 2$. The low accuracy of the estimate of σ^2 here is because we discretized the Gaussian process by 5 dimensional multivariate normal. When using higher dimensional multivariate normal to approximate Gaussian process, the accuracy of the estimate of σ^2 is getting better.

5.2 Simulation Results

Tables 5.3 and 5.4 provide the mean of K-L loss, MSE($\hat{\sigma}^2$) and the mean of $\hat{\sigma}^2$ across 100 simulated samples under symmetric and skewed cases respectively. The mean of K-L loss and MSE($\hat{\sigma}^2$) decrease when number of subjects increase. The mean of $\hat{\sigma}^2$ suggests that $\hat{\sigma}^2$ underestimate σ^2. In addition, the mean of $\hat{\sigma}^2$ is getting closer to the true σ^2 when the number of subjects is getting larger. Also, the mean of K-L loss and MSE($\hat{\sigma}^2$) are getting smaller if the number of bins

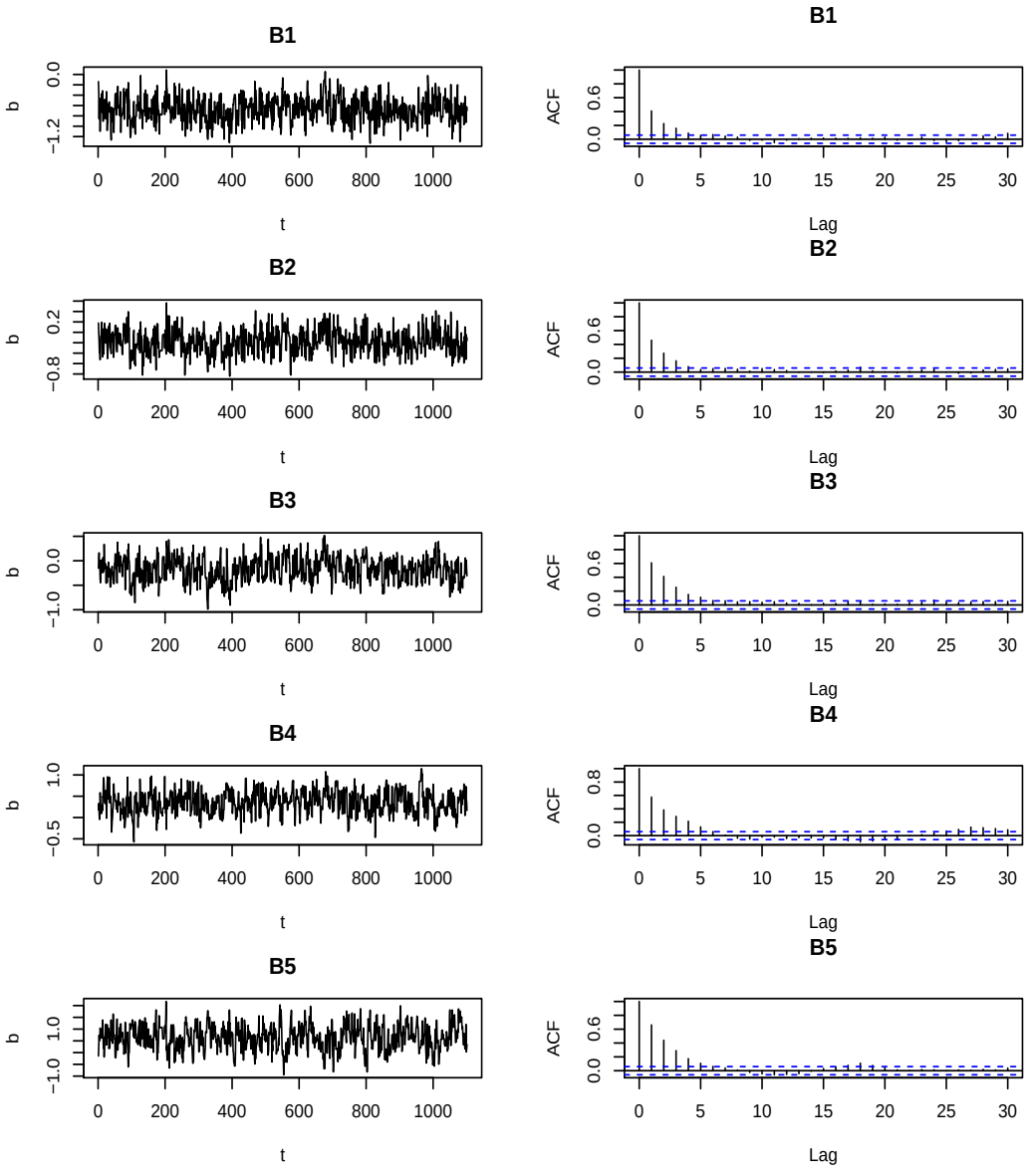


Figure 5.2: Sample MCMC results for $\sigma^2 = 3$: Trace (left column) and ACF (right column) plots based on 1000 effective samples with $\theta = 0.25$, $q = 5, a = 0.38$.

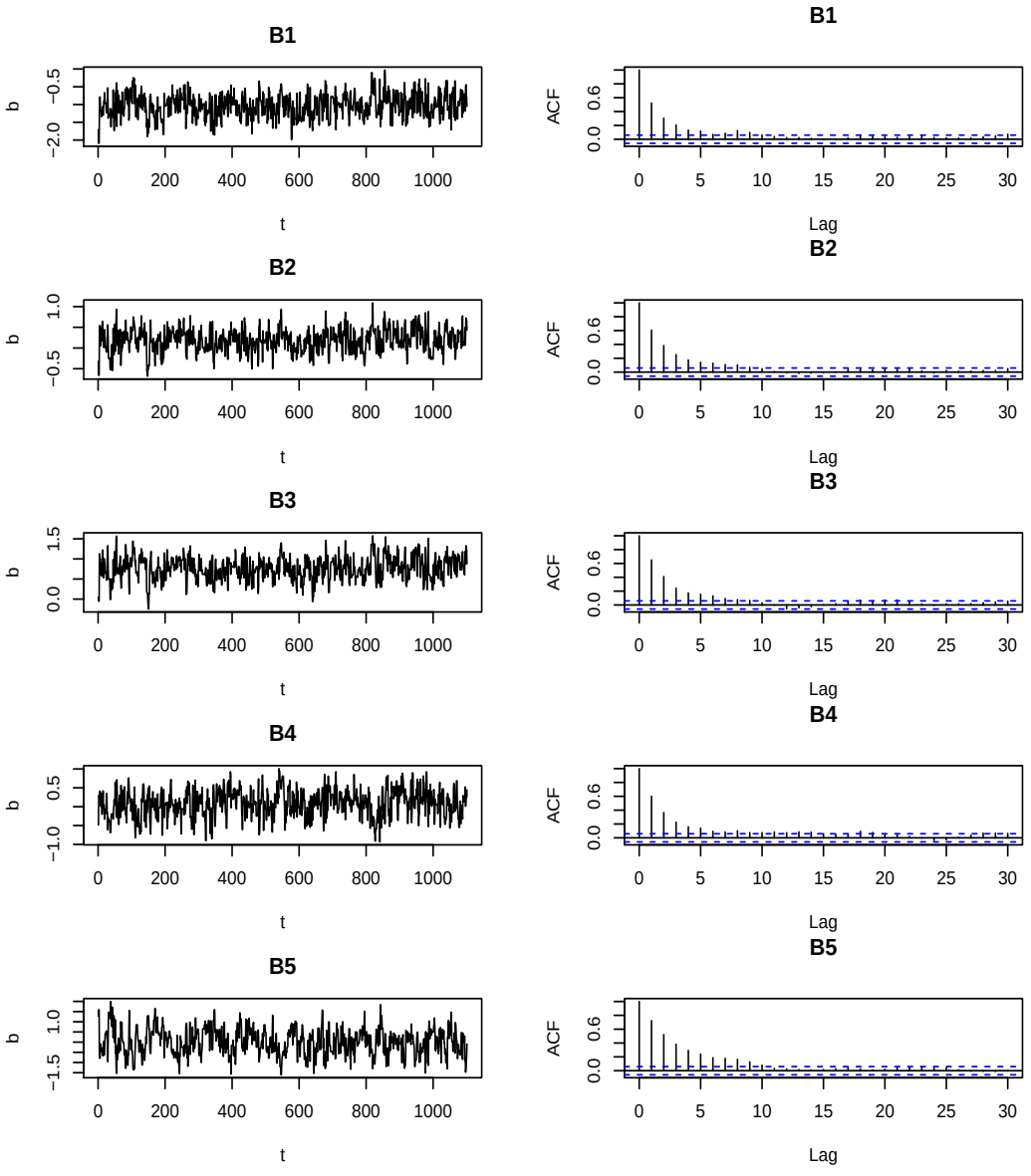


Figure 5.3: Sample MCMC results for $\sigma^2 = 2$: Trace (left column) and ACF (right column) plots based on 1000 effective samples with $\theta = 0.25$, $q = 5, a = 0.38$.

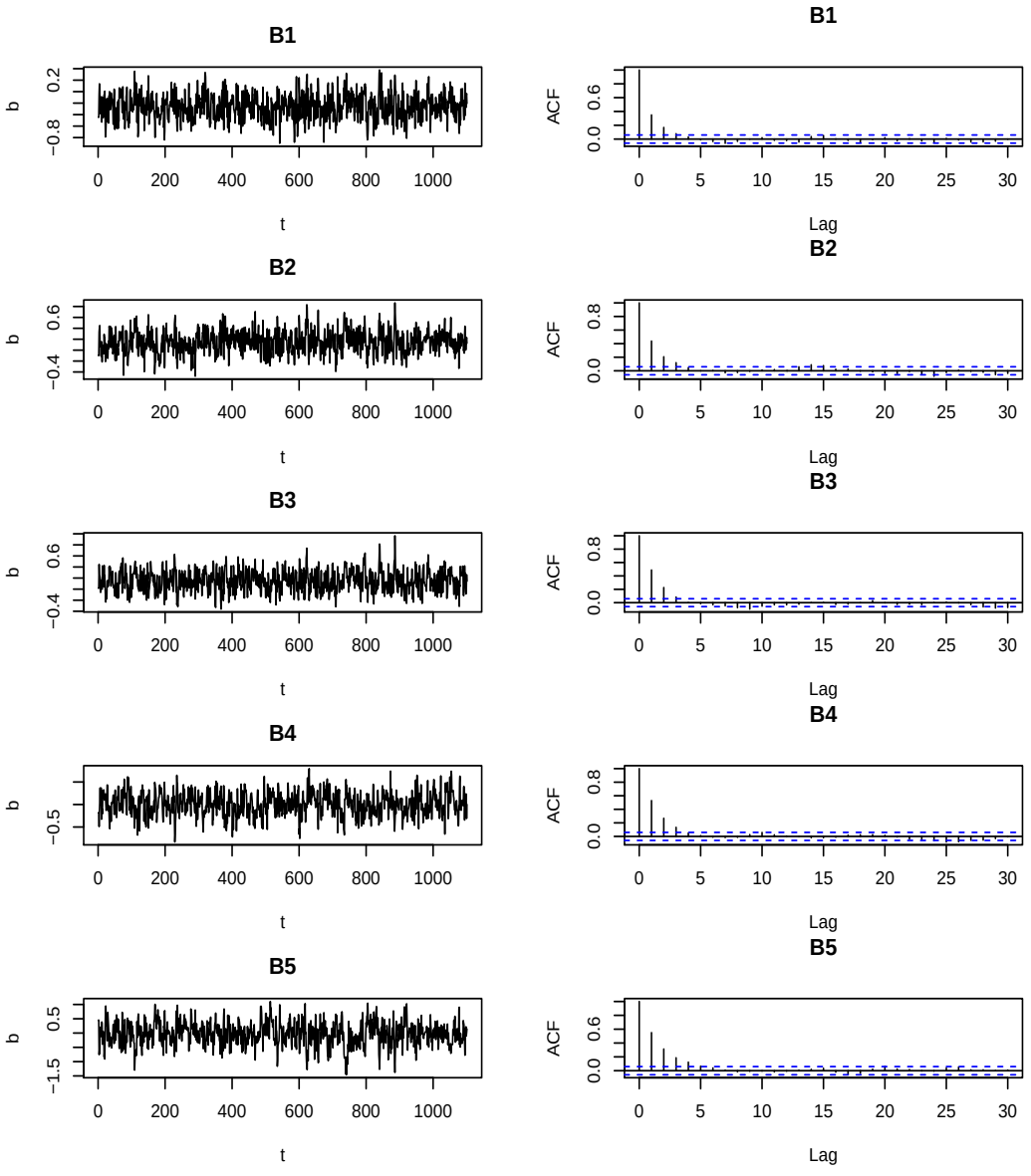


Figure 5.4: Sample MCMC results for $\sigma^2 = 1$: Trace (left column) and ACF (right column) plots based on 1000 effective samples with $\theta = 0.25$, $q = 5, a = 0.38$.

$\theta = 0.5$			$m = 13$	$m = 30$	$m = 100$
$\sigma^2 = 0.5$	$q = 5$	mean of $KL(f, \hat{f})$	0.0053	0.0042	0.0034
		$MSE(\hat{\sigma}^2)$	0.0605	0.0467	0.0379
		mean of $\hat{\sigma}^2$	0.2826	0.2952	0.3095
$\sigma^2 = 0.5$	$q = 10$	mean of $KL(f, \hat{f})$	0.0037	0.0017	0.0007
		$MSE(\hat{\sigma}^2)$	0.0403	0.0230	0.0160
		mean of $\hat{\sigma}^2$	0.3325	0.3750	0.3839
$\sigma^2 = 2$	$q = 5$	mean of $KL(f, \hat{f})$	0.0089	0.0061	0.0039
		$MSE(\hat{\sigma}^2)$	0.9667	0.7746	0.6743
		mean of $\hat{\sigma}^2$	1.0694	1.1509	1.1871
$\sigma^2 = 2$	$q = 10$	mean of $KL(f, \hat{f})$	0.0087	0.0044	0.0017
		$MSE(\hat{\sigma}^2)$	0.6356	0.4987	0.3987
		mean of $\hat{\sigma}^2$	1.2575	1.3279	1.3774

Table 5.3: Performance under the symmetric case ($\theta = 1/2$).

is bigger which means that the estimator geeting closer to the true model and paramter if the raw data are binned with larger number of bins. This is expected, since more bins means more information from raw data are kept and being used to estimate the mode.

Figures 5.5 and 5.6 indicate that for fixed number of subjects, the mean of K-L loss is higher when the value of true variance parameter σ^2 is bigger. Figure 5.7 and 5.8 show how the number of subjects influence mean of KL loss and MES($\hat{\sigma}^2$). Figures 5.9 and 5.10 show the population densities plot for true (f.true), NMDR estimate (f.est), pooled estimate (f.pool) and mean (f.mean) densities for particular simulated data in symmetric and skew case. The mean density estimate is calculated by averaging the smoothing spline density estimates for each single subject. Figures 5.11 and 5.12 are example result plots of true subject densities with their predictions for symmetric and skew population densities respectively.

$\theta = 0.25$			$m = 13$	$m = 30$	$m = 100$
$\sigma^2 = 0.5$	$q = 5$	mean of $KL(f, \hat{f})$	0.0034	0.0027	0.0018
		$MSE(\hat{\sigma}^2)$	0.0642	0.0496	0.0398
		mean of $\hat{\sigma}^2$	0.2768	0.2903	0.3058
$\sigma^2 = 0.5$	$q = 10$	mean of $KL(f, \hat{f})$	0.0030	0.0014	0.0005
		$MSE(\hat{\sigma}^2)$	0.0360	0.0286	0.0146
		mean of $\hat{\sigma}^2$	0.3516	0.3609	0.3867
$\sigma^2 = 2$	$q = 5$	mean of $KL(f, \hat{f})$	0.0069	0.0034	0.0022
		$MSE(\hat{\sigma}^2)$	1.0793	0.8487	0.6763
		mean of $\hat{\sigma}^2$	1.0082	1.1062	1.1891
$\sigma^2 = 2$	$q = 10$	mean of $KL(f, \hat{f})$	0.0071	0.0032	0.0011
		$MSE(\hat{\sigma}^2)$	0.6092	0.5184	0.4054
		mean of $\hat{\sigma}^2$	1.2799	1.3120	1.3750

Table 5.4: Performance under the skewed case ($\theta = 1/4$).

They are both produced with 13 subjects and $\sigma^2 = 2$. The estimate of the subject-specific density, the subject conditional density in the form of the NMDR model, is defined as follows,

$$\hat{f}(y, \hat{b}_i) = \frac{e^{\hat{\eta}_1 + \hat{b}_i(y)}}{\int_0^1 e^{\hat{\eta}_1(y) + \hat{b}_i(y)} dy},$$

where the realization $\hat{b}_i = \{\hat{b}_i(y) | y \in [0, 1]\} = \tilde{E}(\mathbf{B}|\mathbf{Y}_i)$ and $\tilde{E}(\mathbf{B}|\mathbf{Y}_i)$ is approximated by MCMC samples.

5.2.1 The ivestigation of GM estimate of σ^2

Denote $\widehat{\sigma}^2_{GM}$ as GM estimate of σ^2. When estimating σ^2, one can consider taking $\widehat{\sigma}^2_{GM}$ as the initial value for the updating procedure in (4.31) with ζ being replaced by σ^2. In this section, we conduct some simulations to study the accuracy of $\widehat{\sigma}^2_{GM}$. The true parameter σ^2 is set to be 2. Denote m* as the number of subjects, q* as the number of bins and n* as the number of observations generated

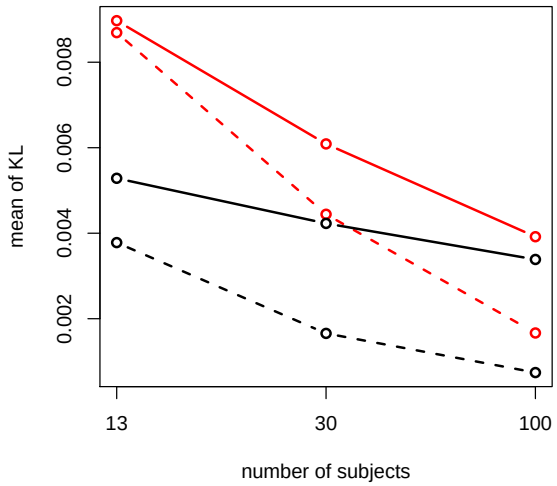


Figure 5.5: Mean of K-L loss for symmetric case ($\theta = 1/2$): black: $\sigma^2 = 0.5$, red: $\sigma^2 = 2$; solid: $q = 5$, dotted: $q = 10$.

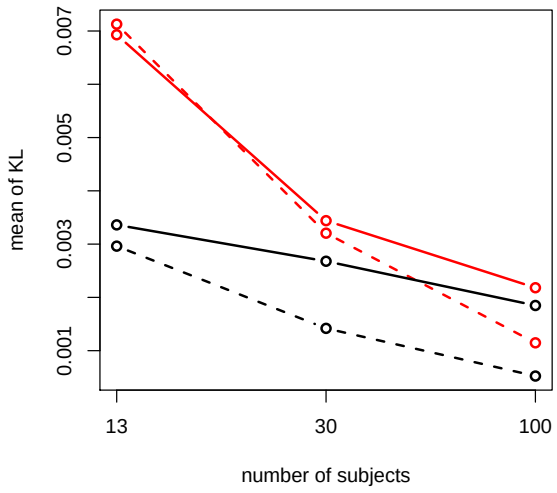


Figure 5.6: Mean of K-L loss for skewed case ($\theta = 1/4$): black: $\sigma^2 = 0.5$, red: $\sigma^2 = 2$; solid: $q = 5$, dotted: $q = 10$.

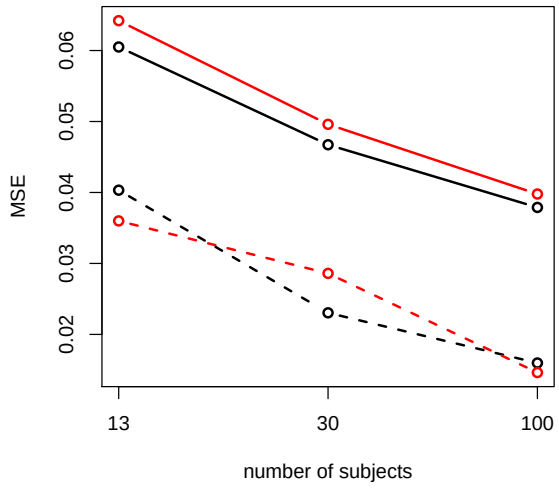


Figure 5.7: MSE of $\hat{\sigma}^2$, for $\sigma^2 = 0.5$: black: symmetric ($\theta = 1/2$), red: skew ($\theta = 1/4$); solid: $q = 5$, dotted: $q = 10$.

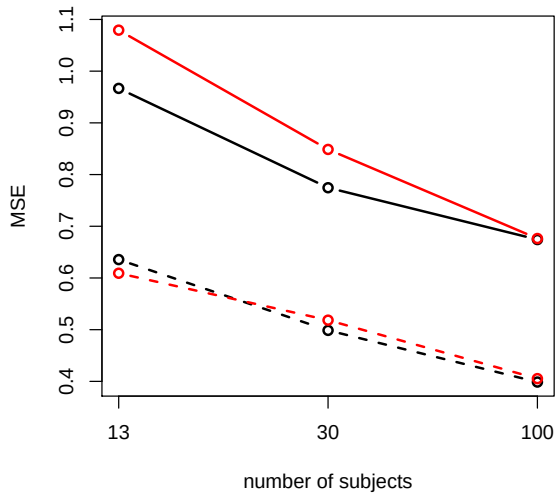


Figure 5.8: MSE of $\hat{\sigma}^2$, for $\sigma^2 = 2$: black: symmetric ($\theta = 1/2$), red: skew ($\theta = 1/4$); solid: $q = 5$, dotted: $q = 10$.

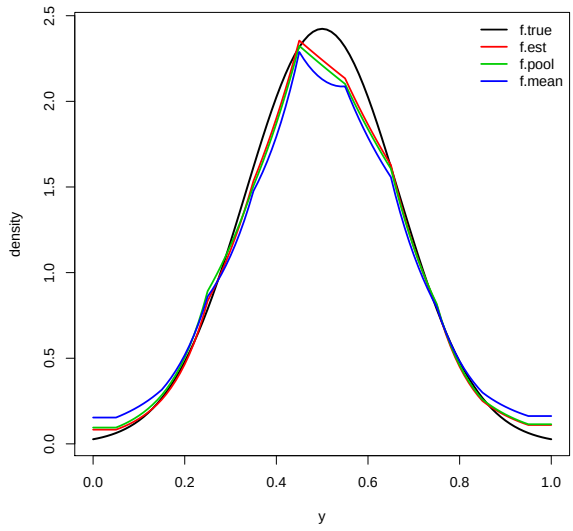


Figure 5.9: Plots of the true density and its estimates for the symmetric case ($\theta = 1/2$) based on a particular simulated sample.

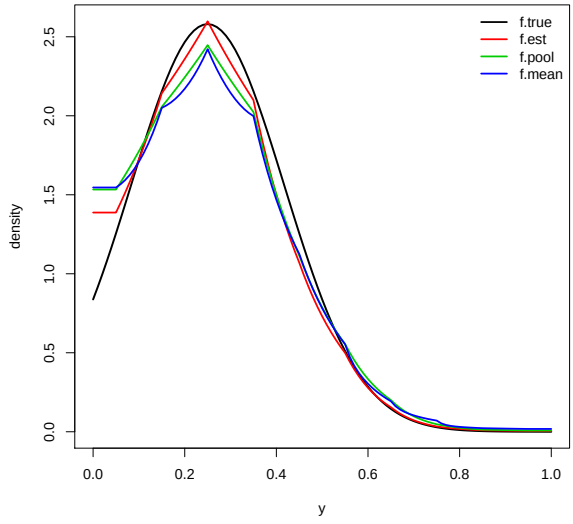


Figure 5.10: Plots of the true density and its estimates for the skewed case ($\theta = 1/4$) based on a particular simulated sample.

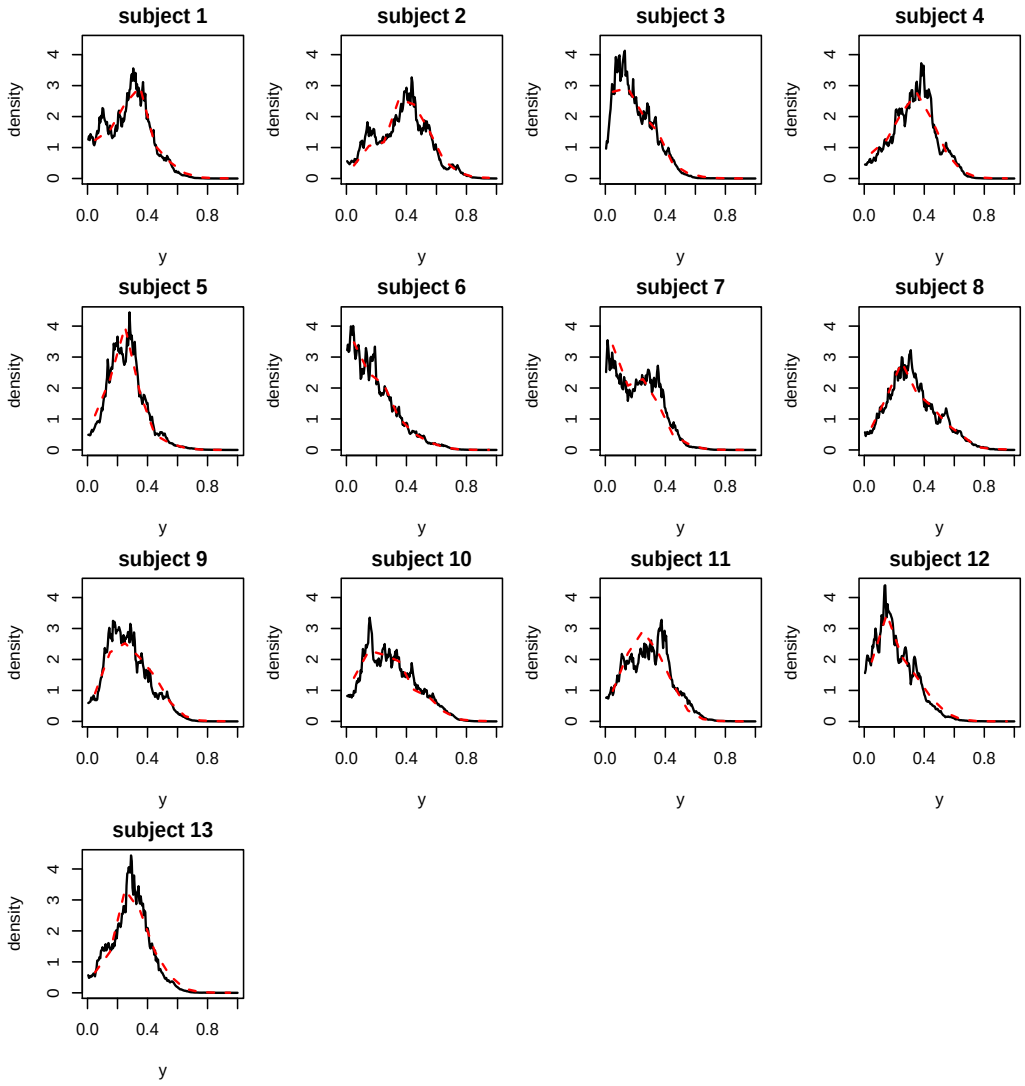


Figure 5.11: Plots of true subject-specific densities (black solid lines) and their estimate (red dotted lines) for symmetric case ($\theta = 1/2$) based on a particular simulated sample.

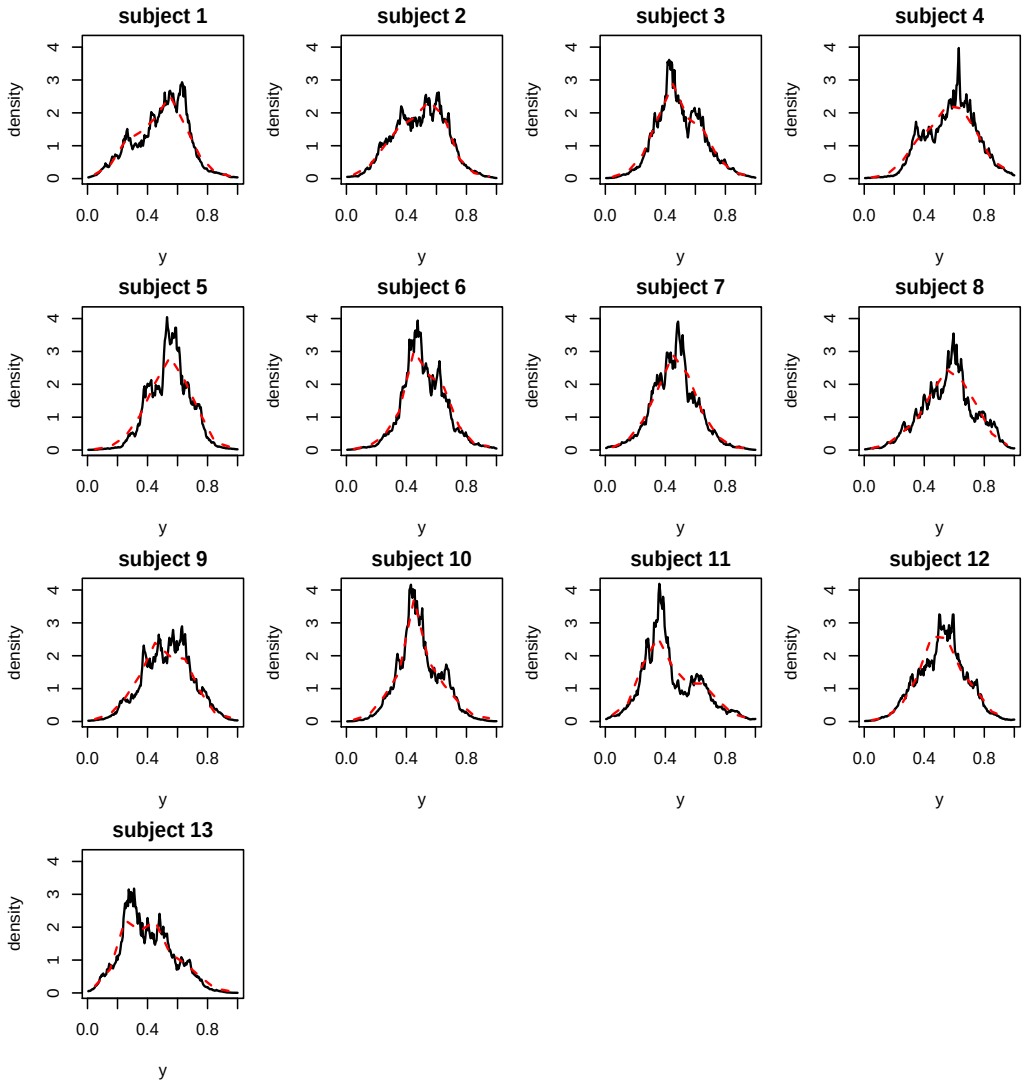


Figure 5.12: Plots of true subject-specific densities (black solid lines) and their estimate (red dotted lines) for skewed case ($\theta = 1/4$) based on a particular simulated sample.

from each subject. We run 100 experiments for each of two different shapes of population density (symmetric and skewed), m* (30 and 100), q* (10 and 20) and n* (200 and 1000). Figures 5.13 and 5.14 display the boxplots of $\log(\widehat{\sigma}^2_{GM}/\sigma^2)$. The setting, (m*,q*,n*), for each experiment is listed as follows, experiment 1: (30,10,200), experiment 2: (30,10,1000), experiment 3: (30,20,200), experiment 4: (30,20,1000), experiment 5: (100,10,200), experiment 6: (100,10,1000), experiment 7: (100,20,200) and experiment 8: (100,20,1000). We can see that with more bins and bigger number of subjects and number of observations in each subject, the GM estimate is closer to the true parameter which means that the GM estimate might be a good initial value under these conditions.

5.3 Smoothness comparison

In this section we compare the smoothness of linear and cubic spline estimates. The cubic spline NMDR model approximates the main effect by $\eta_1(y) = d \times (y - 0.5) + \sum_{l=1}^{q} c_l R_2(K_l, y)$ and models the covariance of random effects as $\sigma_1^2 \times (s - 0.5)(t - 0.5) + \sigma_2^2 R_2(s,t)$ where $R_2(s,t) = k_2(s)k_2(t) - k_4(|s-t|)$.

We generate data from the model $f(y, b_i) = e^{\eta_1(y)+b_i(y)} / \int_0^1 e^{\eta_1(y)+b_i(y)} dy$ where $\eta_1(y) = -3(y - 1/2)^2$ and $b_i = \{b_i(y) | y \in [0,1]\}$ is a realization of a Gaussian process with mean 0 and covariance function $50R_2(s,t)$. Again, the domain $[0,1]$ is discretized by dividing it into 200 subregions for generating raw data. The

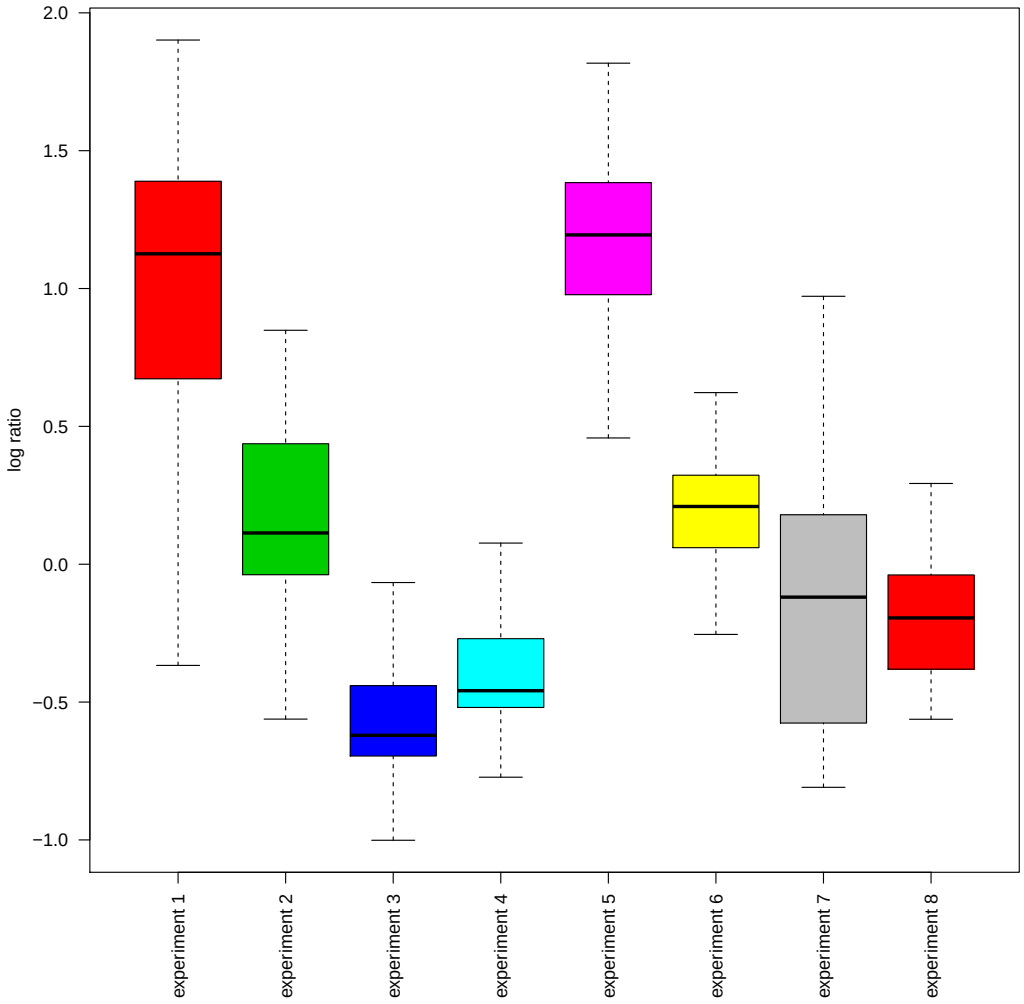


Figure 5.13: Log ratio boxplots for symmetric case ($\theta = 1/2$). The vertical axis represents the log ratio $\log(\widehat{\sigma}^2_{GM}/\sigma^2)$.

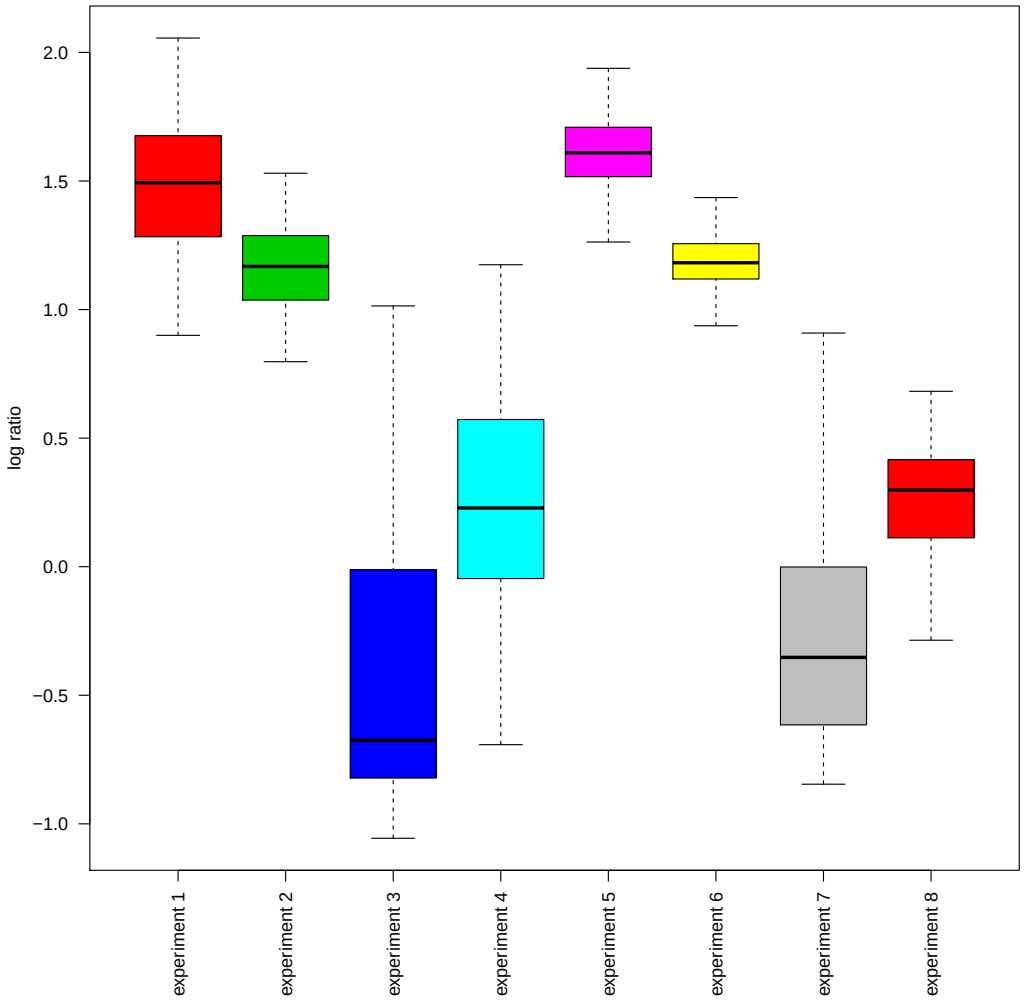


Figure 5.14: Log ratio boxplots for skewed case ($\theta = 1/4$). The vertical axis represents the log ratio $\log(\widehat{\sigma}^2_{GM}/\sigma^2)$.

simulated data are then grouped into 20 equal length bins for each subjects. The number of subjects is set to be 30 and each subject has 200 observations.

Figures 5.15-5.16 display the comparison of the true population density with its linear and cubic estimates. Figures 5.17-5.20 display the comparison of true subject density and its linear and cubic predictions. We only run one experiment for each shape of density to illustrate the smoothness of linear and cubic estimates. The cubic spline estimates and predictions are smoother than linear spline. One may also notice that the true subject densities in figures 5.15 and 5.16 are smoother than those in 5.11 and 5.12, this is because the model we use to generate data in this section used a cubic kernel as covariance function while in the previous section about large scale simulation is with linear kernel. And models with cubic covariance function are smoother than linear.

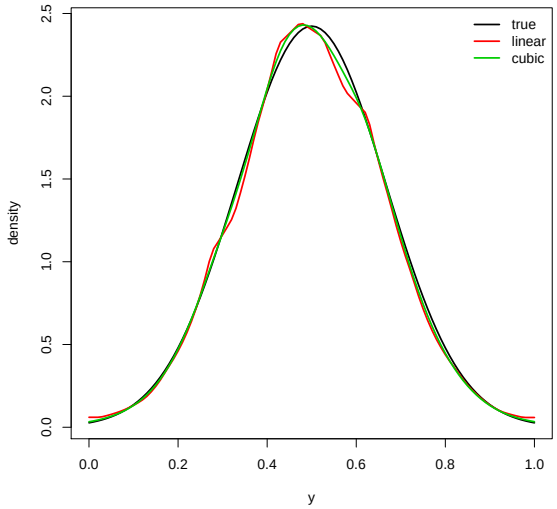


Figure 5.15: Population density estimates comparison for symmetric case based on a particular sample.

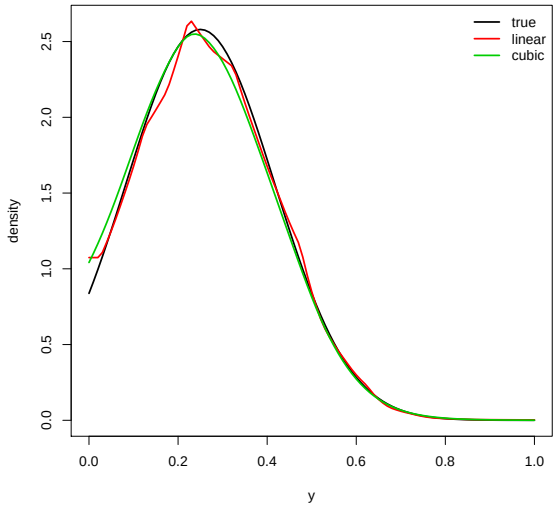


Figure 5.16: Population density estimates comparison for skewed case based on a particular sample.

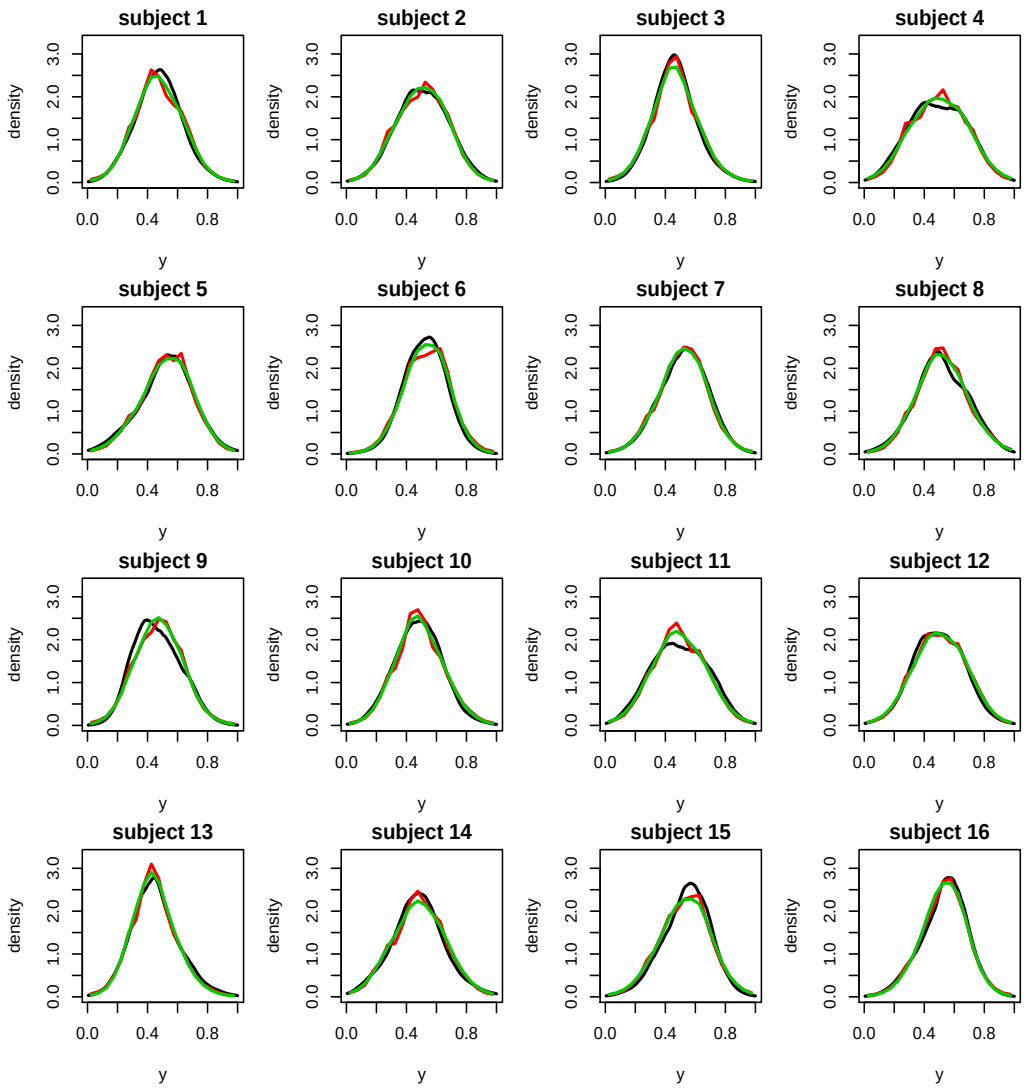


Figure 5.17: Subject-specific density and its estimates: symmetric case, subject 1-16. Black: true subject-specific densities; Red: linear spline estimates; Green: cubic spline estimates.

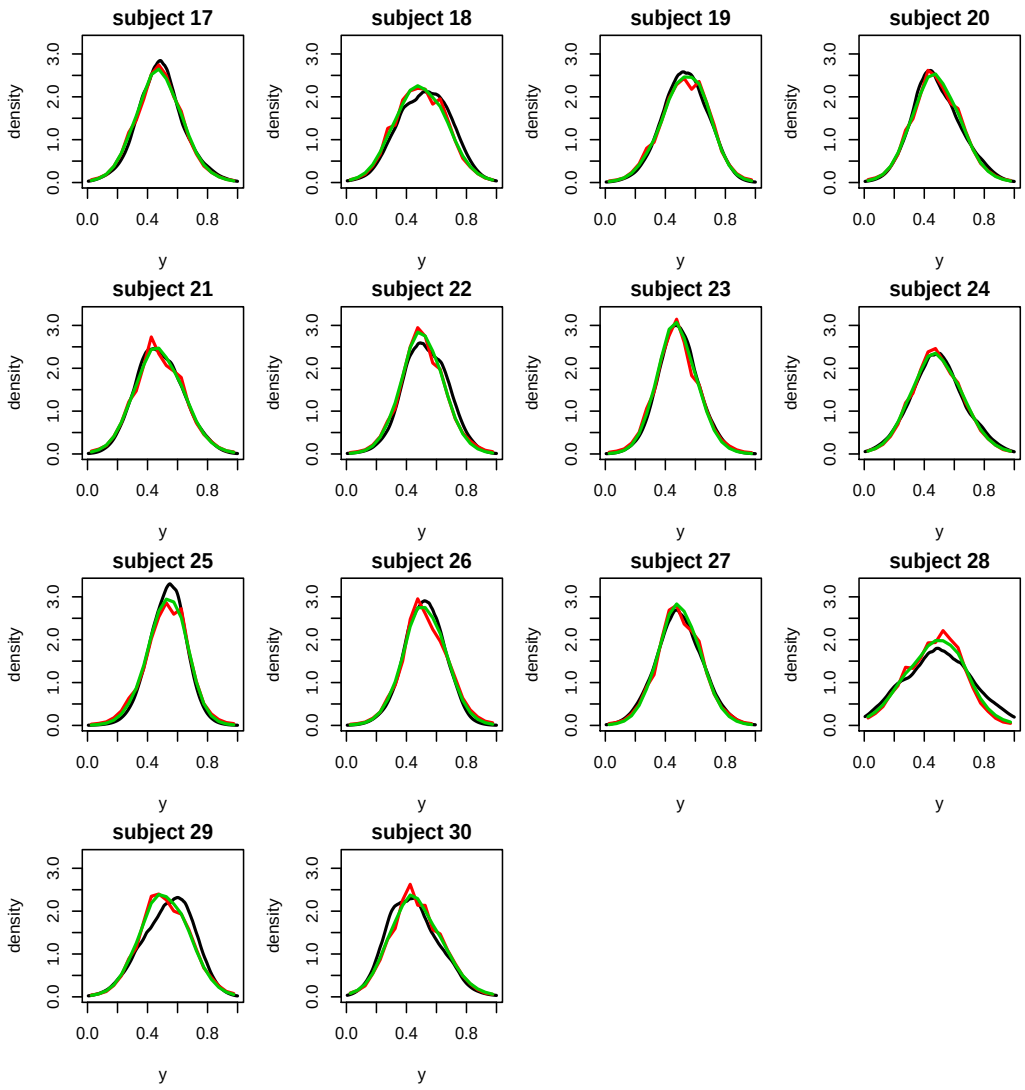


Figure 5.18: Subject-specific density and its estimates: symmetric case, subject 17-30. Black: true subject-specific densities; Red: linear spline estimates; Green: cubic spline estimates.

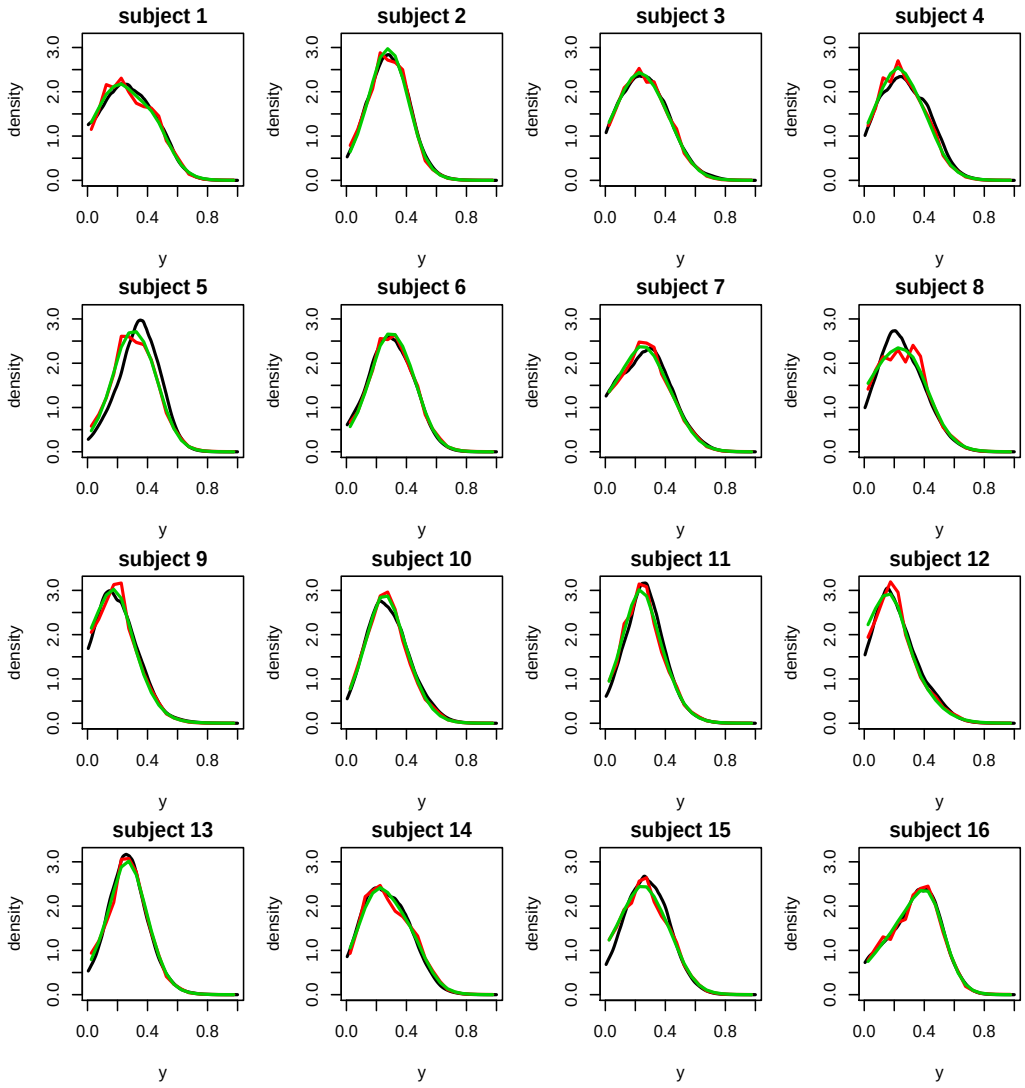


Figure 5.19: Subject density and its estimates: skewed case, subject 1-16. Black: true subject-specific densities; Red: linear spline estimates; Green: cubic spline estimates.

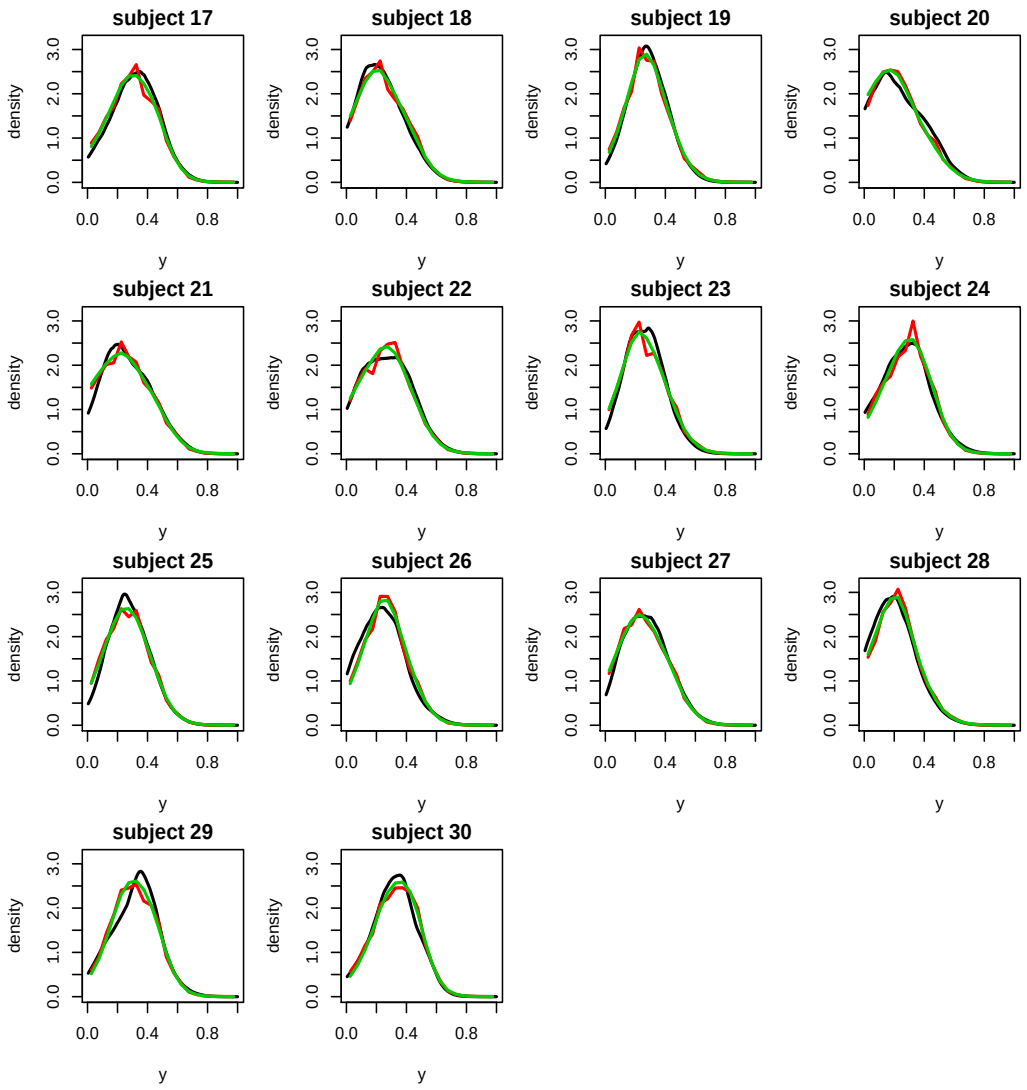


Figure 5.20: Subject density and its estimates: skewed case, subject 17-30. Black: true subject-specific densities; Red: linear spline estimates; Green: cubic spline estimates.

Chapter 6

Application to Speech Data

6.1 Scientific Questions and Data

The objective in this chapter is to compare the phonation interval (PI) distributions (especially in the short PI region: 30–150-ms) between normal speakers and people who stutter during oral reading. We also compare the difference during the stutter-free speech (i.e., when recorded intervals of speech containing stuttering were removed ; Godinho et al., 2006).

According to the website of National Stuttering Association (NSA), stuttering is a communication disorder involving disruptions, or "disfluencies," in a person's speech. Gow and Ingham (1992) found that a reduction in short phonated intervals (PIs) in the range of 30-150-ms is associated with decreased stuttering. Ingham et al. (2001) showed that purposefully reducing the number of short PIs resulted in the elimination of stuttering.

The PI intervals can be viewed as an estimate of the duration of vocal fold movement. For instance, a 50-ms PI refers to a 50-ms period during which the

vocal folds were vibrating (Ingham et al., 2001). Speakers produce a number of PIs of varying duration in a specified amount of speaking time (Davidow, Bothe, Andreatta a nd Ye 2009). Additional information for the PI measurement can be found in Davidow et al., 2009.

In our dataset the experiment involved 13 adult and adolescent individuals who stuttered and 13 control participants who were matched for age and gender. The domain, 30–1000-ms range, was subdivided into 50-ms ranges (except for the 30–50-ms range, which was left as a 20-ms subdivision) and the total number of PIs from each subject that occurred within each of these 20 subdivisions provided the raw data (Godinho et al., 2006). Figure 6.1 shows the nonparametric density estimator of PI distribution for each of two datasets: the data set contains all time-periods (top row) and the dataset after all stuttering periods removed (bottom row). The R package `gss` (Gu 2009) is used to estimate density function for each participant. Visually, there is a large variation between subjects.

Our dataset is from Godinho et al. (2006). They apply t-test to detect the difference in the proportion of PIs of each subdivision between two groups. Their finding suggests no difference in the distribution of PIs between normal subjects group and people who stutter group. However, using a t-test for the problem of interest may lose information contained in smooth density functions since a t-test is a test based on using the means as summaries instead of a test based on the entire density for each individual and group. Our NMDR model takes the smoothness of the density function into account during the density estimation, and

hence keep more information from data when detecting the difference between the two groups. The result based on our proposed method supports their finding.

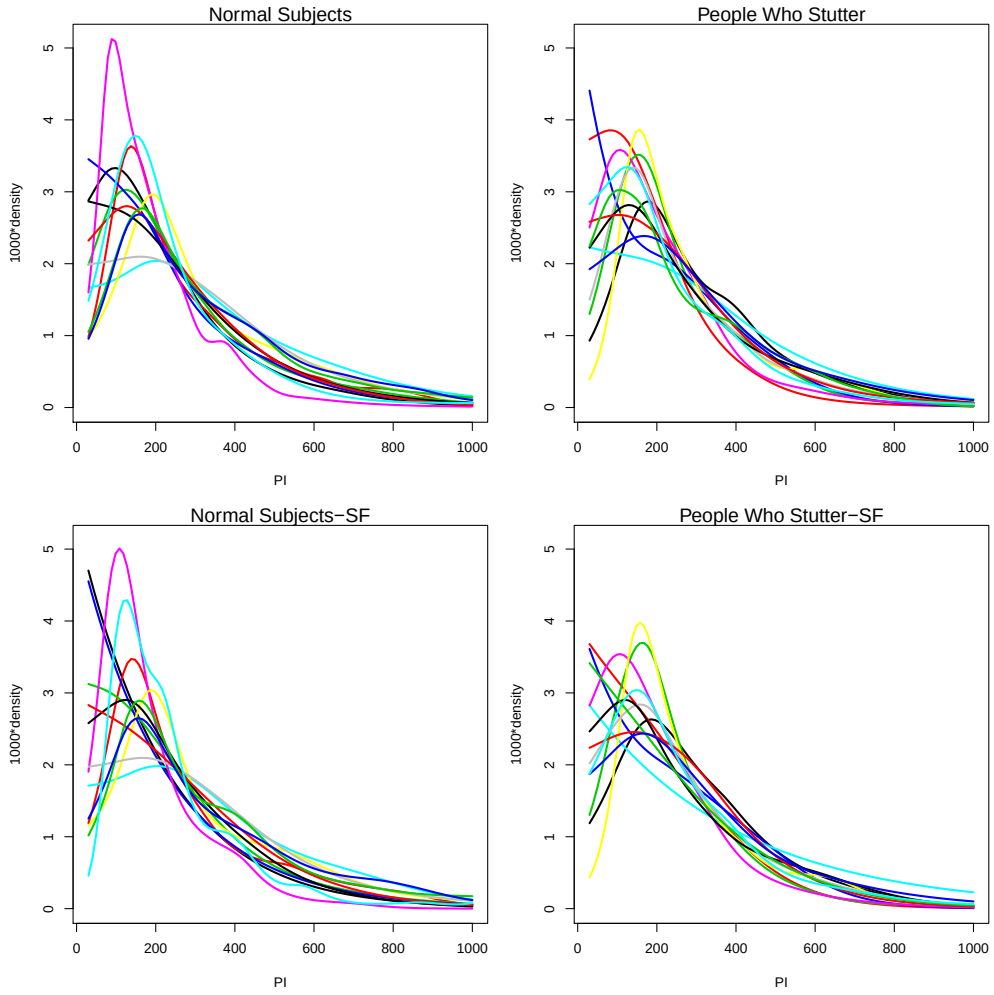


Figure 6.1: Density estimations for two different groups (normal speaker/stutter) under two different datasets (complete: top row; stutter-free speech: bottom row). Different colors represent different subjects.

6.2 Initial Analysis

To compare the PI distributions in the region 30–150-ms, we prepare two plots to display the distribution of PI odd. We compute the odd of short PI by

$$\text{odd} = \frac{\text{proportion of P-I from 30–150-ms}}{\text{proportion of P-I from 150–1000-ms}}.$$

The odd tells us the ratio of proportions between short PIs (30–150-ms) and long PIs (150–1000-ms). Small odd indicates the subject speaks with more long PIs. Figure 6.2 contains boxplots of PI odd for complete and stutter-free dataset. They suggest the odd of the group of normal subjects and the group of people who stutter are similar.

Since the data are paired we also check the ratio of odds of normal subject to stutter subject. The odds ratio for the i^{th} pair of subjects is computed by

$$OR_i = \frac{\text{odd of normal subject } i}{\text{odd of stutter subject } i}.$$

Figure 6.3 displays the distribution of odds ratio. In Figure 6.3, we see that for complete dataset most odds ratio are below one which indicates normal people tend to use less short intervals than people who stutter. For stutter-free dataset, the median is closed to one.

We use logistic regression to test the difference in odd between normal speakers and people who stutter separately for each of the two datasets. In this case, we do not consider data as paired. Here we define the odds ratio as the odds of normal people divided by the odds of people who stutter. The estimates for the

log odds ratio of the short region are -0.0359 ($p-value = 0.157$) and -0.0206 ($p-value = 0.559$) for complete dataset and stutter-free dataset respectively. Neither is significant at 5% significance level which implies the differences are not significant in both datasets at 5% significance level. In addition, we use mixed-effects logistic regression to test the difference in odds between two groups when data are considered as paired for two datasets. The estimates for the log odds ratio of the short intervals are -0.0626 ($p-value = 0.015$) and 0.0358 ($p-value = 0.334$) for the complete dataset and stutter-free dataset respectively. Hence, at 5% significance level, the difference in odds between is significant in complete dataset when data are considered as paired.

6.3 Fitting NMDR Models

In this section we use NMDR model developed in Chapter 3 to estimate the population densities and subject-specific densities. We fit model for each group separately. In each dataset, each group is estimated by linear (3.4) and cubic (3.4) spline NMDR models. We write the NMDR model for subject-specific density as,

$$f(y, b_i) = \frac{e^{\eta_1(y)+b_i(y)}}{\int_{\mathcal{Y}} e^{\eta_1(y)+b_i(y)}dt}.$$

where $b_i = \{b_i(y)|y \in \mathcal{Y}\}$ and $\mathcal{Y} = [30, 1000]$.

For the linear spline case (3.4), $b_i's$ are realizations of independent Gaussian processes with mean 0 and covariance function $\sigma^2 R_1(s,t)$, where R_1 is linear spline kernel. For the cubic spline case (3.6), the functional random effect has two

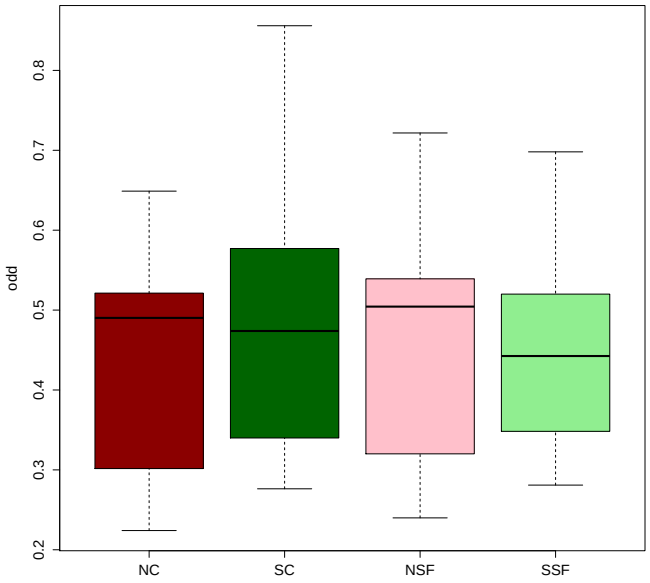


Figure 6.2: Boxplots of odds for normal and stutter subjects from both datasets. Dark red: Normal speakers from complete dataset (NC). Dark green: Stutterer from complete dataset (SC). Pink: Normal speakers from stutter-free dataset (NSF). Light green: Stutterer from stutter-free dataset (SSF).

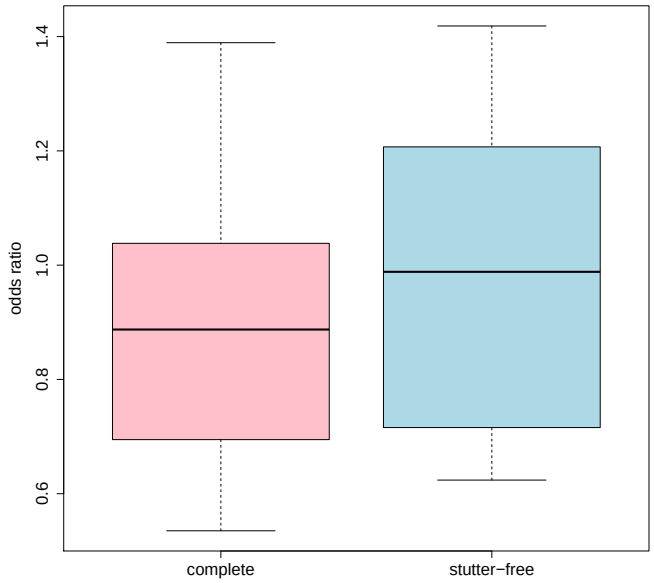


Figure 6.3: Boxplots of odd ratios for complete (pink) and stutter-free (blue).

components, $b_i(y) = b_{i,1} \times (y - 0.5) + b_{i,2}(y)$ where $b'_{i,1}s$ are realizations of i.i.d random variables from $N(0, \sigma_1^2)$ and $\{b_{i,2}(y)|y \in \mathcal{Y}\}$, $i = 1, ..., n$, are realizations of independent Gaussian processes with mean 0 and covariance function $\sigma_2^2 R_2(s, t)$ where R_2 is cubic spline kernel. Hence in the cubic spline case, the functional random effect b_i is a realization of Gaussian process with mean 0 and covariance function $\sigma_1^2 \times (s - 0.5)(t - 0.5) + \sigma_2^2 R_2(s, t)$.

The interval $\mathcal{Y} = [30, 1000]$ is divided into 20 subdivisions where the first bin has length 20ms and all others are 50ms long. We use the middle point of each subdivision to be a knot. Hence the number of knots $L = 20$. The parameter *alpha* in the cross-validation score in (4.28) for smoothing parameter selection is set to be 1 for the linear spline case and 1.4 for the cubic spline case.

We write t_j, $j = 1, ..., 20$ as the middle points of each bin. Let R_1 and R_2 be the linear and cubic spline kernels respectively. Denote $\Sigma_l^{(k)}$ as a 20 by 20 matrix with $(i,j)^{th}$ element $\sigma^{2(k)} R_1(t_i, t_j)$. Let $\Sigma_c^{(k)}$ be a 21 by 21 matrix with $(i,j)^{th}$ element $\sigma_2^{2(k)} R_2(t_i, t_j)$ if $1 \leq i, j \leq 20$ and $\sigma_1^{2(k)}$ if $i = j = 21$, off-diagonal elements in the 21^{th} row and column are set equal to zero. In linear spline case, the Metropolis-Hastings proposal distribution at the k^{th} iteration is 20 dimensional $MVN(0, a^2\Sigma_l^{(k)})$. In cubic spline case, the functional random effect has two mutually independent components $b_{i,1}$ and $b_{i,2}(t)$. Write $\mathbf{b}_i = (b_{i,2}(t_1), ..., b_{i,2}(t_{20}), b_{i,1})^T$. At the k^{th} iteration, we simulate $\mathbf{b}_i$ from 21 dimensional $MVN(\mathbf{0}, a^2\Sigma_c^{(k)})$.

The value of a is a tuning parameter chosen to keep the acceptance rates around 23%. Based on several simulation studies that follows the procedure described in section 5.1.3, we decide to use $a = 0.36$ and 0.34 for linear and cubic case respectively.

We store every 10^{th} MCMC sample after an initial burn-in of 200 sweeps. The maximum number of iterations is 150for the whole updating procedure. The updating procedure stops when $\frac{||\zeta^{(k)} - \zeta^{(k-1)}||}{||\zeta^{(k-1)}||} < 5 \times 10^{-4}$. We use SAA with the step and MCMC size at the k^{th} step to be $\gamma_k = 1$ and $m_k = m_0 + k^2$. m_0 is set to be 200 and 2500 for updating $(\mathbf{c}, \mathbf{d})$ and ζ respectively. Note $\zeta = \sigma^2$ and $\zeta = (\sigma_1^2, \sigma_2^2)$ for linear and cubic case.

The initial value of $(\mathbf{c}, \mathbf{d})$ is set to be pooled estimate, $(\mathbf{c}^{(0)}, \mathbf{d}^{(0)}) = (\widehat{\mathbf{c}}_{pooled}, \widehat{\mathbf{d}}_{pooled})$. For the variance parameters, we use large value for initial value, $\zeta^{(0)} = \sigma^{2(0)} = 2$ and $\zeta^{(0)} = (\sigma_1^{2(0)}, \sigma_2^{2(0)}) = (1, 50)$ for the linear and cubic spline model, respectively.

6.4 Results

We estimate the subject-specific density $f(y, b_i)$ by

$$\hat{f}(y, b_i) = \frac{e^{\hat{\eta}_1(y)+\hat{b}_i(y)}}{\int_{30}^{1000} e^{\hat{\eta}_1(y)+\hat{b}_i(y)} dy},$$

where $\hat{\eta}_1$ can be obtained by *ssden* in R library **gss** and $\hat{b}_i(y) = \tilde{E}(\mathbf{B}_i|Y = y)$ is computed by MCMC sampling.

Figures 6.4 (linear spline estimate) and 6.5 (cubic spline estimate) indicate that population and subject density estimates for each group are skewed to the right. Figures 6.6 and 6.7 compare the NMDR estimates with pooled estimates. The pooled estimates, directly combining data across subjects, has higher peak than linear spline NMDR estimates, however it is almost identical to the cubic spline NMDR estimates.

Figures 6.8 to 6.9 are population density comparisons (using NMDR estimates) between normal people and people who stutter. The black vertical dotted line represents the boundary between short and long PI. The plots do not suggest significant difference in the short PI region between normal people and people who stutter. Figures 6.10 to 6.12 are NMDR estimate for each subject densities. Again, the black vertical dotted line represents the boundary between short and long PI.

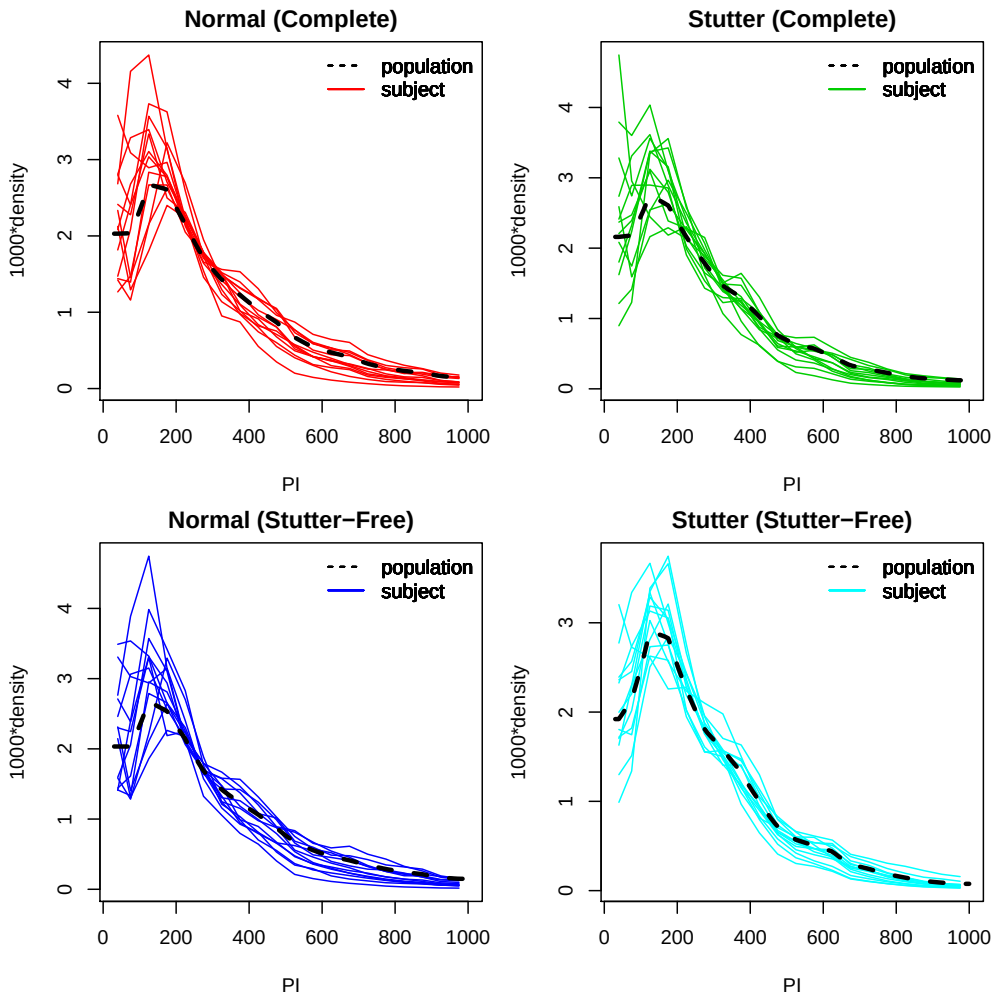


Figure 6.4: Linear spline estimates of population and subject-specific density functions: The first row are plots for the complete dataset. The second row are plots for the stutter-free dataset.

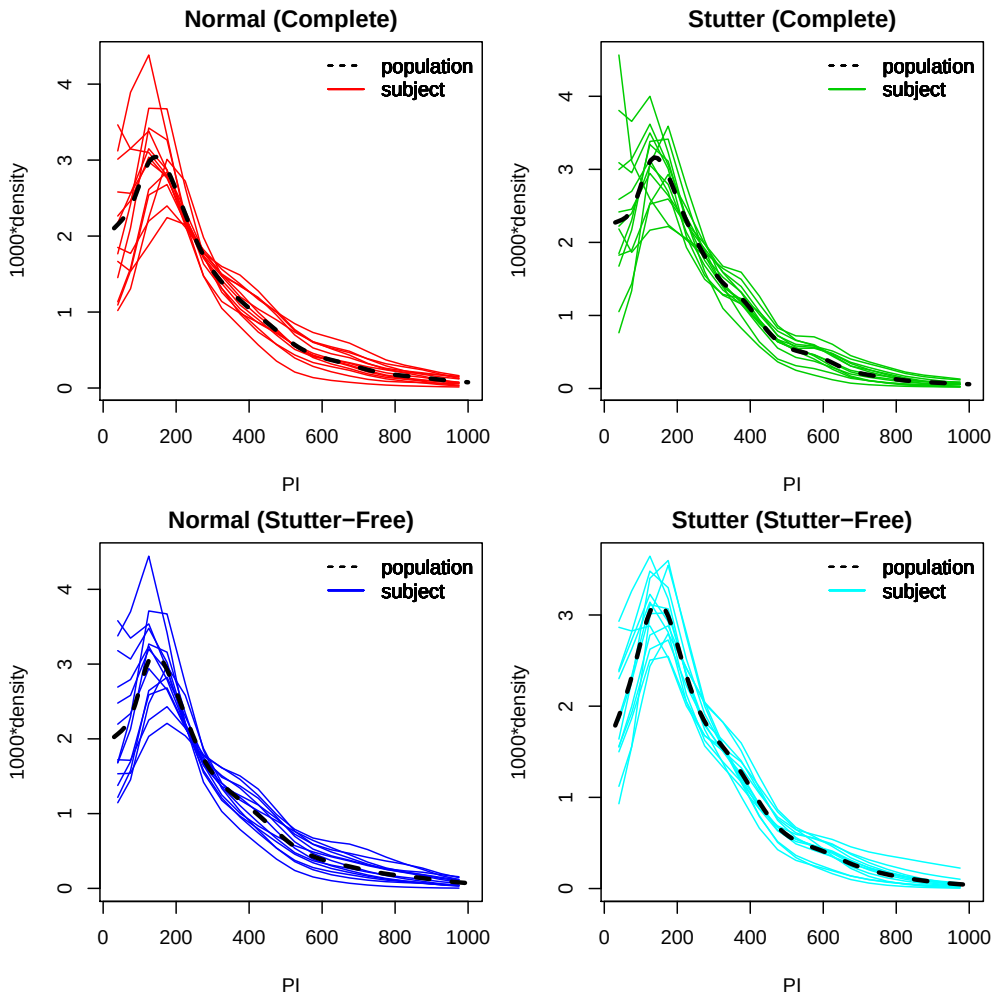


Figure 6.5: Cubic spline estimates of population and subject-specific density functions: The first row are plots for the complete dataset. The second row are plots for the stutter-free dataset.

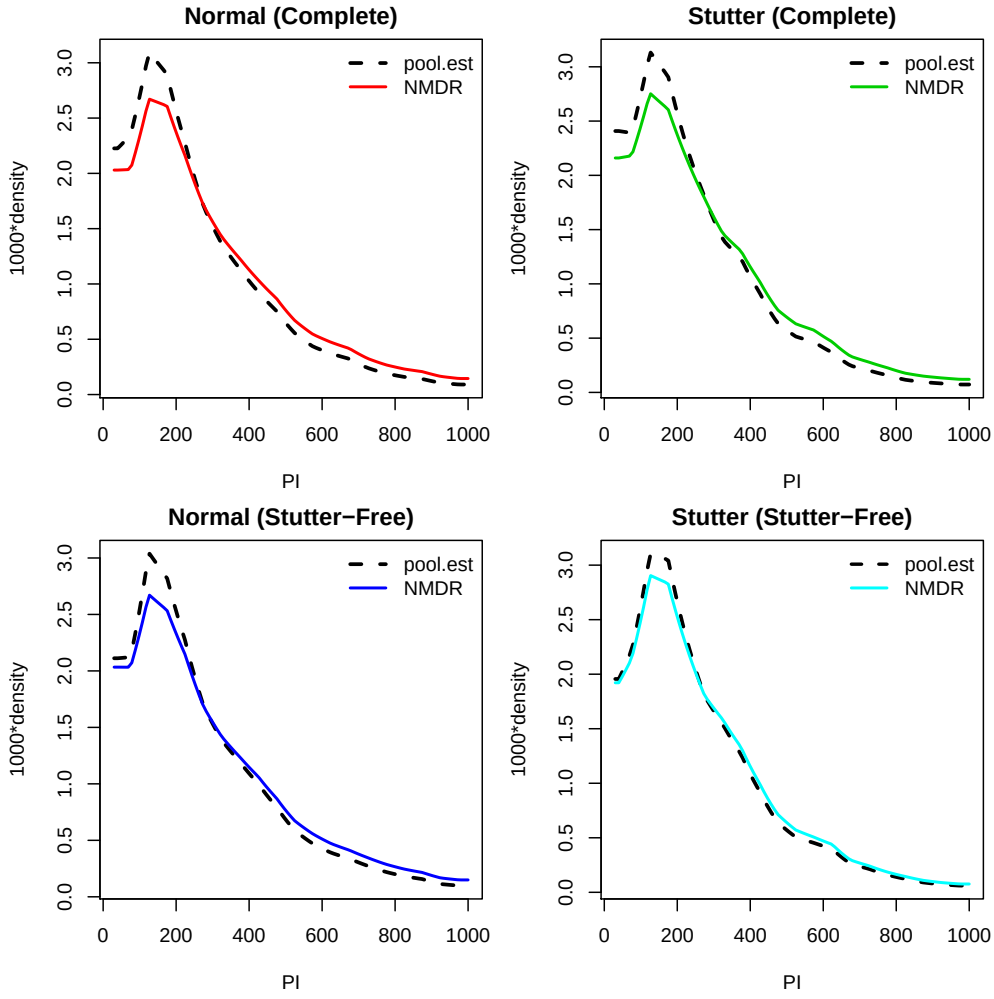


Figure 6.6: Linear spline population densities estimates: The first row are plots for the complete dataset. The second row are plots for the stutter-free dataset.

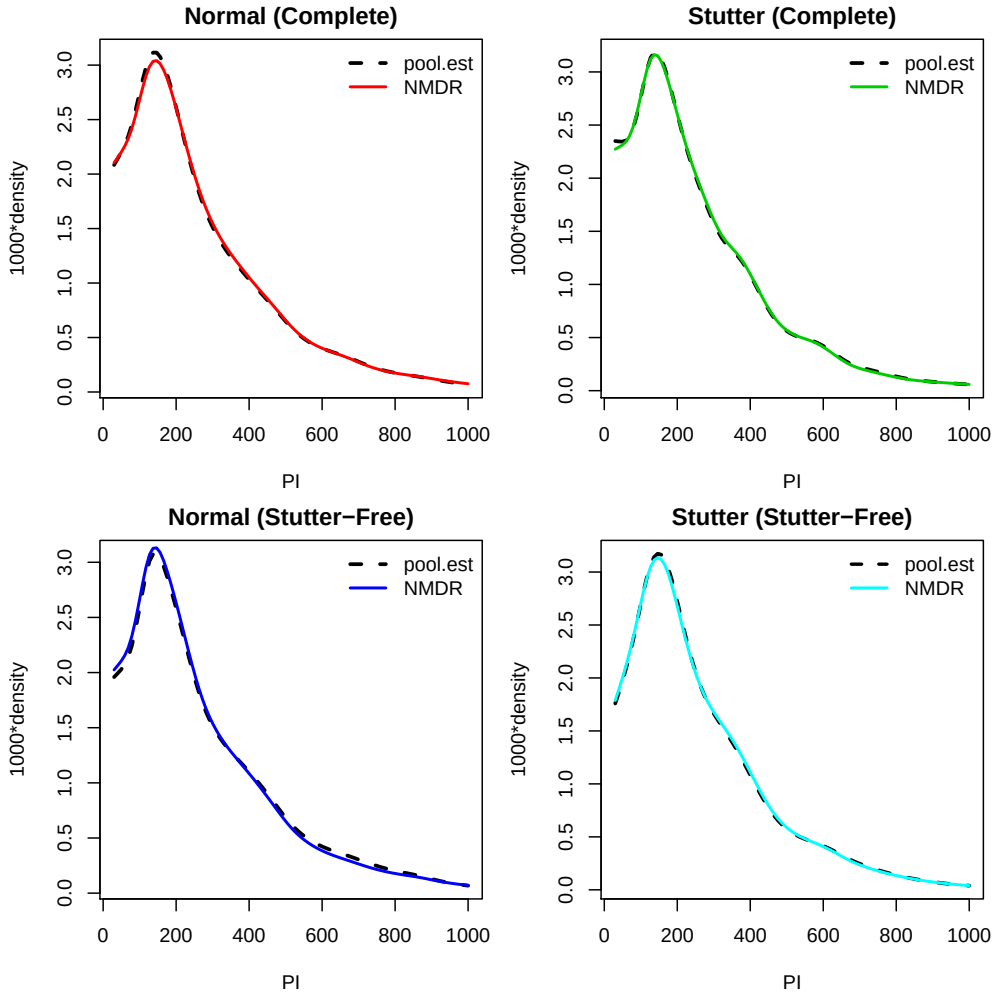


Figure 6.7: Cubic spline population densities estimates: The first row are plots for the complete dataset. The second row are plots for the stuter-free dataset.

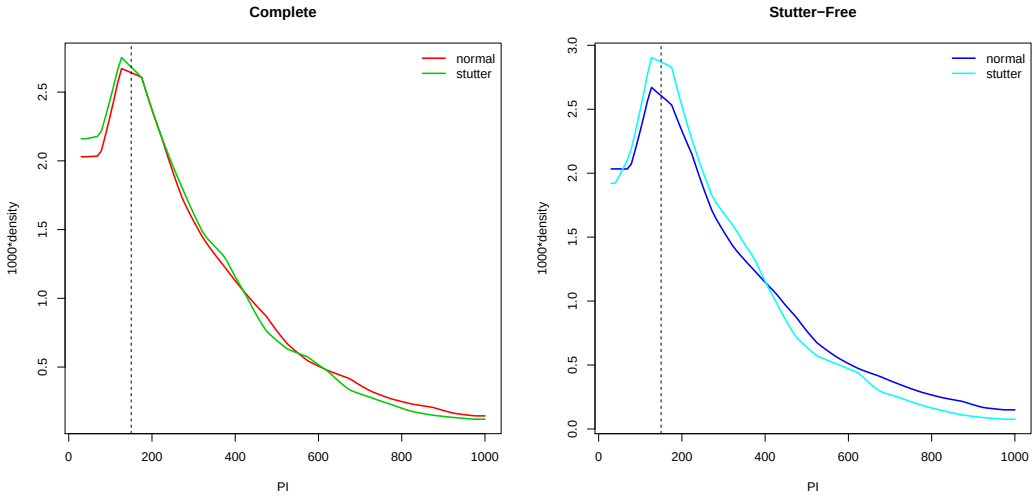


Figure 6.8: Linear spline population density estimates plots: Complete dataset (left), Stutter-Free dataset (right).

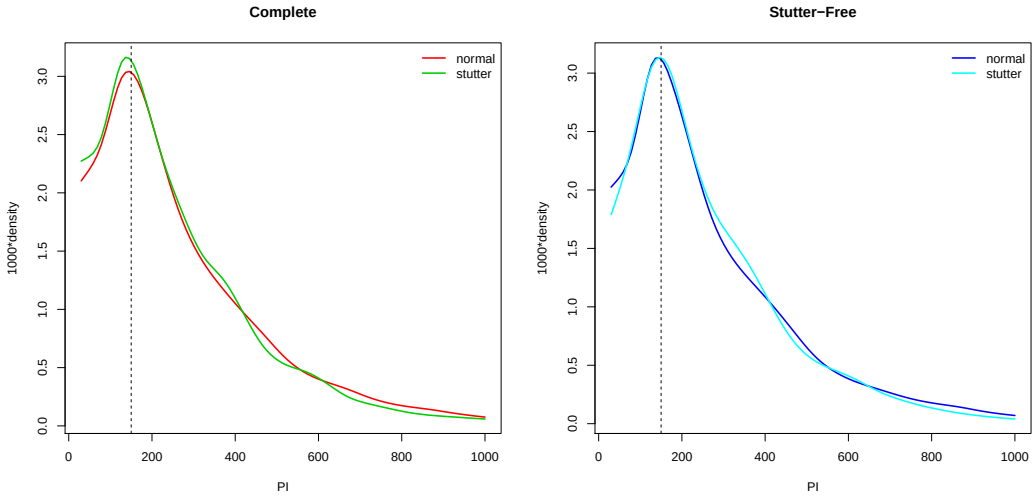


Figure 6.9: Cubic spline population density estimates plots: Complete dataset (left), Stutter-Free dataset (right).

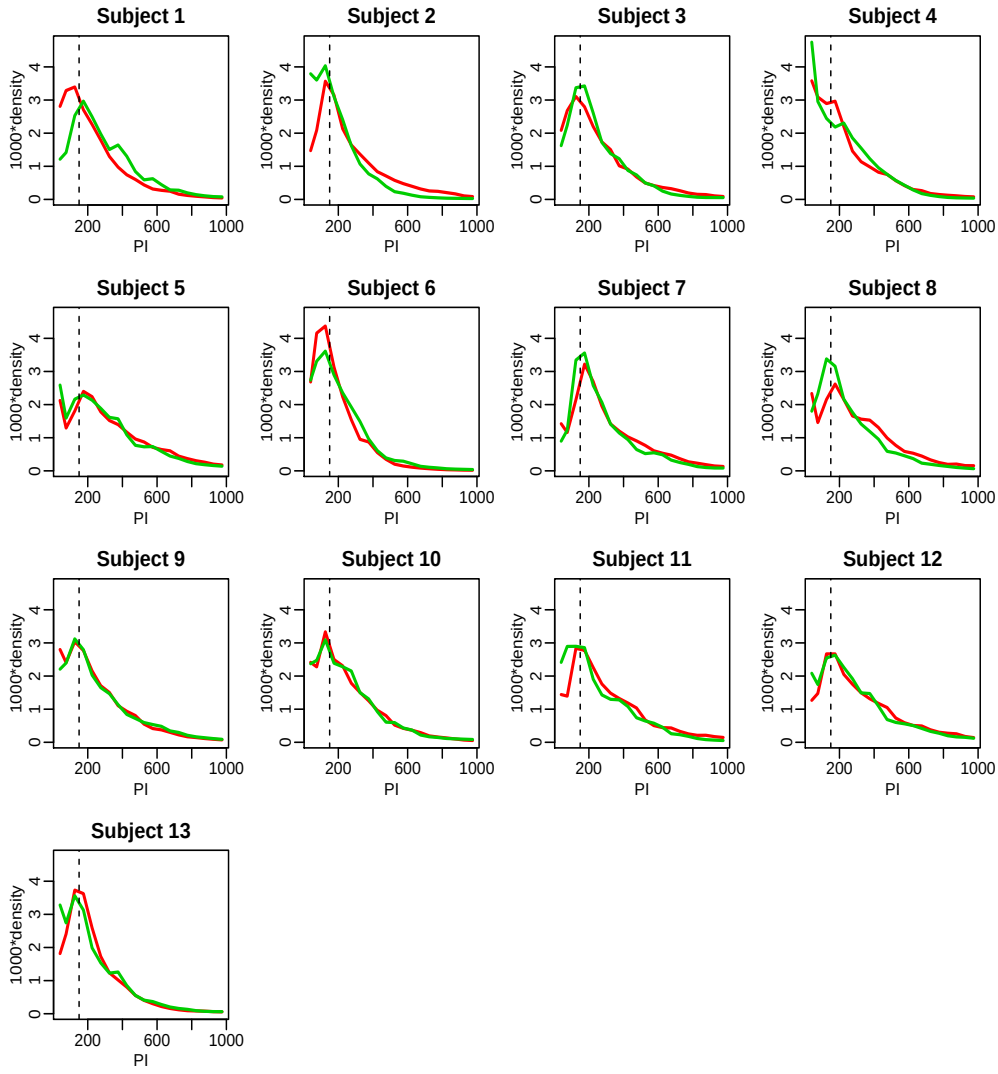


Figure 6.10: Linear spline subject-specific density estimates for the complete dataset. Red: Normal Subject. Green: Stutter Subject.

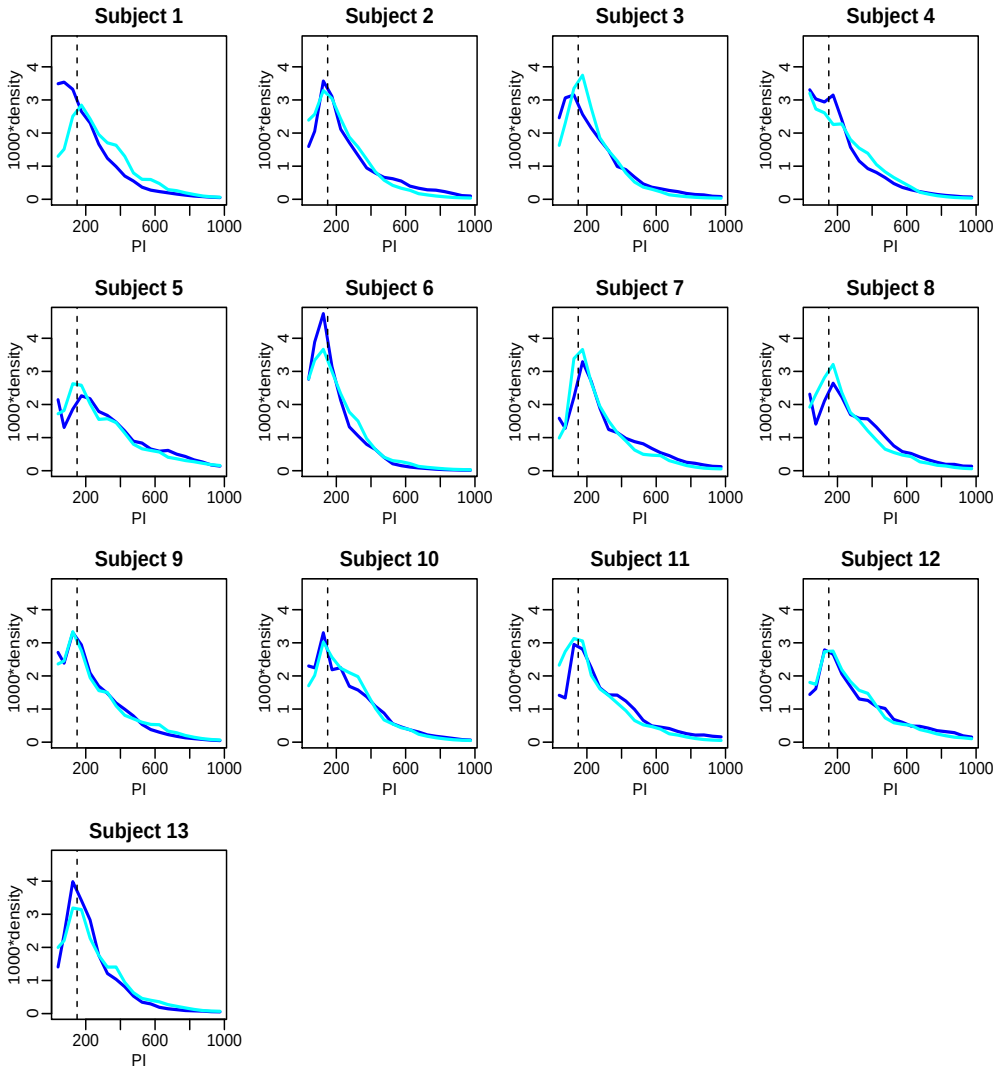


Figure 6.11: Linear spline subject-specific density estimates for the stutter-free dataset. Blue: Normal Subject. Cyan: Stutter Subject.

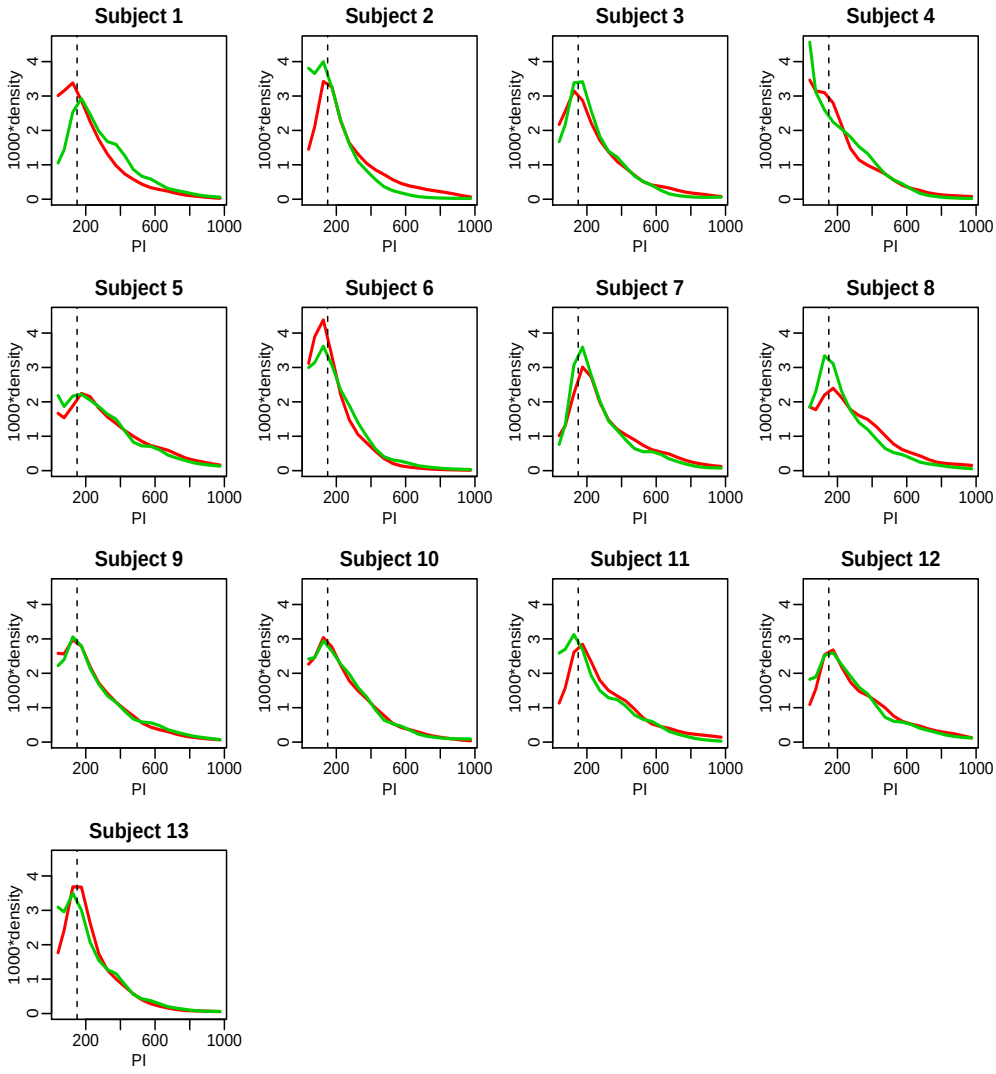


Figure 6.12: Cubic spline subject-specific density estimates for the complete dataset. Red: Normal Subject. Green: Stutter Subject.

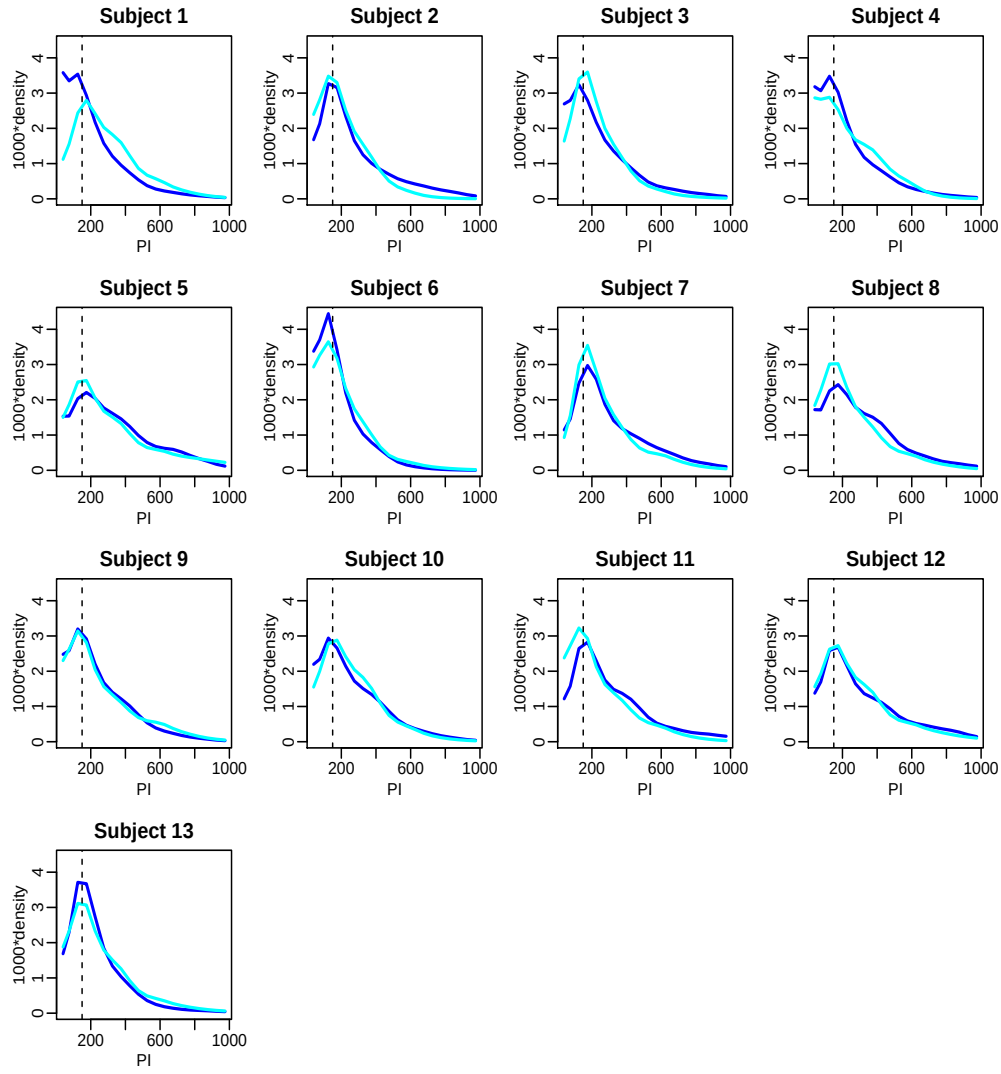


Figure 6.13: Cubic spline subject-specific density estimates for the stutter-free dataset. Blue: Normal Subject. Cyan: Stutter Subject.

Dataset	Group	$\hat{A}_S$	$\hat{A}_L$	$\hat{A}_C$
Complete	Normal	0.3048	$0.2760[0.2582, 0.2991]$	$0.3106[0.2968, 0.3309]$
Complete	Stutter	0.3212	$0.2901[0.2822, 0.3184]$	$0.3234[0.3143, 0.3448]$
Stutter-Free	Normal	0.3123	$0.2761[0.2406, 0.3057]$	$0.3066[0.2938, 0.3514]$
Stutter-Free	Stutter	0.3063	$0.2910[0.2789, 0.3309]$	$0.3067[0.2879, 0.3300]$

Table 6.1: The area estimates of short PI region.

6.4.1 Comparison in the Area of Short PI Region

The goal of this analysis is to compare the area of short PI regions between normal speakers and people who stutter. In this section, we provide the comparison among different estimates when the paired effect are not taken into the consideration. Denote $A = \int_{30}^{150} f(y)dy$ as the the area of short PI regions under the population density curve f. Let $\hat{A}_L$ and $\hat{A}_C$ represent estimates of A based on linear and cubic NMDR population density estimates respectively. Also, denote $\hat{A}_S$ as the estimate computed as the average of sample proportions of short region across 13 subjects, $\sum_{i=1}^{13}\sum_{j=1}^{3}\pi_{ij}/13$ where π_{ij} represents the observed proportion of counts in the j^{th} subinterval for the i^{th} subject. Table 6.1 shows each estimate of the area of the short PI region. The 95% bootstrap confidence intervals based on 50 simulations are also provided for linear and cubic NMDR estimates. The bootstrap approach will be described later in this section.

Table 6.2 shows the various estimates of difference in the area of short PI region. The second column represents the estimate of the difference in the area of short PI region based on sample proportions $\pi'_{ij}s$ between normal speaker $\hat{A}_{SN}$ and people who stutter $\hat{A}_{SS}$. The third column represents the linear NMDR estimates

Dataset	$\hat{A}_{SN} - \hat{A}_{SS}$	$\hat{A}_{LN} - \hat{A}_{LS}$	$\hat{A}_{CN} - \hat{A}_{CS}$
Complete	-0.0164	-0.0141	-0.0128
		B.I=$[-0.0387, 0.0094]$	B.I=$[-0.2828, 0.0015]$
Stutter-Free	0.0060	-0.0149	-0.0005
		B.I=$[-0.0523, 0.0165]$	B.I=$[-0.0157, 0.0468]$

Table 6.2: The estimates of difference in the area of short PI region between the two groups. $\hat{A}_{SN}$, $\hat{A}_{LN}$ and $\hat{A}_{CN}$ are estimates based on sample proportions, linear and cubic NMDR models respectively for normal speakers. $\hat{A}_{SS}$, $\hat{A}_{LS}$ and $\hat{A}_{CS}$ are estimates for people who stutter.

of the difference in the area of short PI region between normal speaker $\hat{A}_{LN}$ and people who stutter $\hat{A}_{LS}$. The last column represents the cubic NMDR estimates of the difference in the area of short PI region between normal speaker $\hat{A}_{CN}$ and people who stutter $\hat{A}_{CS}$. In addition, the 95% bootstrap confidence intervals based on 50 simulations are provided for linear and cubic NMDR estimates. The 95% intervals all suggest no significance between the two groups.

Table 6.3 shows the various estimates of log odds ratio for the area of short PI region between normal speakers and people who stutter. The second column shows the log odds ratio, $log(OR) = log(odd(N)/odd(S))$ where $odd(N) = \hat{A}_{SN}/(1 - \hat{A}_{SN})$ and $odd(S) = \hat{A}_{SS}/(1 - \hat{A}_{SS})$. The third and fourth column display quantities $LORL$ and $LORC$ which are log odds ratios based on linear and cubic NMDR population density estimates respectively. The 95% bootstrap confidence intervals based on 50 simulations are also provided for linear and cubic

Dataset	$log(OR)$	$LORL$	$LORC$
Complete	-0.0763	-0.0694	-0.0591
		B.I=$[-0.1893, 0.0512]$	B.I=$[-0.1304, 0.0089]$
Stutter-Free	0.0281	-0.0733	-0.0002
		B.I=$[-0.2537, 0.0854]$	B.I=$[-0.0748, 0.2297]$

Table 6.3: The estimates of log odds ratio for the area of short PI region between the two groups.

NMDR estimates. The 95% intervals all suggest no significance between the two groups.

The approach we use to construct the confidence intervals is called basic bootstrap confidence limit. Details for this approach can be found in Davision and Hinkley (1997). We will only describe the algorithm briefly. Assume that we have m subjects in each group. Let $\mathbf{x}_i$ and $\mathbf{y}_i$ be the vectors that collects all observations from the i^{th} subject in normal speaker group and people who stutter group respectively. Also, denote the NMDR population density estimates as $\hat{g}$ for normal speakers and $\hat{h}$ for people who stutter. Set $\hat{A} = \int_{30}^{150} \hat{g}(y)dy$ and $\hat{B} = \int_{30}^{150} \hat{h}(y)dy$ as the estimates for the area of short PI region for each group. Also, denote the log odds ratio $LOR = \log\{[\hat{A}/(1-\hat{A})]/[\hat{B}/(1-\hat{B})]\}$. The algorithm is described as follows,

1. For $r = 1, ..., R$,

   (a) Randomly sample m numbers $\{I_1, ..., I_m\}$ with replacement from $\{1, ..., m\}$;

(b) Compute $\hat{g}_r^*$ and $\hat{h}_r^*$ based on observations $\{\mathbf{x}_{I_1}, ..., \mathbf{x}_{I_m}\}$ and $\{\mathbf{y}_{I_1}, ..., \mathbf{y}_{I_m}\}$ respectively;

(c) Compute $\hat{A}_r^*$ and $\hat{B}_r^*$;

(d) Compute $\hat{d}_r^* = \hat{A}_r^* - \hat{B}_r^*$ and $LOR_r^* = \log\{[\hat{A}_r^*/(1-\hat{A}_r^*)]/[\hat{B}_r^*/(1-\hat{B}_r^*)]\}$;

2. Sort each set of bootstrap estimates: $\{\hat{A}_{(1)}^*, ..., \hat{A}_{(R)}^*\}$, $\{\hat{B}_{(1)}^*, ..., \hat{B}_{(R)}^*\}$, $\{\hat{d}_{(1)}^*, ..., \hat{d}_{(R)}^*\}$ and $\{LOR_{(1)}^*, ..., LOR_{(R)}^*\}$;

3. Compute the $(1-\alpha)$ bootstrap confidence intervals:

$$[2\hat{A} - \hat{A}^*_{((R+1)(1-\alpha))}, 2\hat{A} - \hat{A}^*_{((R+1)\alpha)}],$$

$$[2\hat{B} - \hat{B}^*_{((R+1)(1-\alpha))}, 2\hat{B} - \hat{B}^*_{((R+1)\alpha)}],$$

$$[2\hat{d} - \hat{d}^*_{((R+1)(1-\alpha))}, 2\hat{d} - \hat{d}^*_{((R+1)\alpha)}],$$

$$[2LOR - LOR^*_{((R+1)(1-\alpha))}, 2LOR - LOR^*_{((R+1)\alpha)}].$$

For approximating integrals in computing areas $\hat{A}$ and $\hat{B}$, we first divide the domain $\mathcal{Y} = [30, 1000]$ into 97 equal length bins $[30, 40]$, $[40, 50]$, ... and $[990, 1000]$, each with bin width $10ms$. The first 12 bins represents the region of the short PI. Let K_i be the middle point of the i^{th} subinterval, we approximate $\hat{A} \approx \sum_{i=1}^{12} \hat{g}(K_i)/10$ and $\hat{B} \approx \sum_{i=1}^{12} \hat{h}(K_i)/10$.

Appendix A

Derivative of PL

We compute the derivatives of PL in (4.14) and (4.15). When taking derivatives of (4.4) the penalty term is easy to deal with, so we shall only show the work of computing the derivative of the marginal likelihood $l(\zeta, \mathbf{c}, \mathbf{d})$. Denote $\mathcal{B}$ as the range of $\mathbf{b}_i$. First we need to compute the derivative of $\int p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)p_{\mathbf{B}_i}(\mathbf{b}_i)d\mathbf{b}_i$,

$$\begin{aligned}
& \frac{\partial}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \int_{\mathcal{B}} p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)p_{\mathbf{B}_i}(\mathbf{b}_i)d\mathbf{b}_i \\
= & \int_{\mathcal{B}} [\frac{\partial}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)p_{\mathbf{B}_i}(\mathbf{b}_i)]d\mathbf{b}_i \\
= & \int_{\mathcal{B}} \{\frac{\partial}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]\} p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)p_{\mathbf{B}_i}(\mathbf{b}_i)d\mathbf{b}_i \\
= & \{E_{\mathbf{B}_i|\mathbf{Y}_i}\{\frac{\partial \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T}\}\} p_{\mathbf{Y}_i}(\mathbf{Y}_i).
\end{aligned}$$

Thus the first derivative of marginal likelihood $l(\zeta, \mathbf{c}, \mathbf{d})$ is

$$
\begin{aligned}
& \frac{\partial}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} l(\zeta, \mathbf{c}, \mathbf{d}) \\
= & \frac{\partial}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \{\sum_{i=1}^{m} \log \int_{\mathcal{B}} p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) p_{\mathbf{B}_i}(\mathbf{b}_i) d\mathbf{b}_i\} \\
= & \sum_{i=1}^{m} \{\frac{\partial}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \log \int_{\mathcal{B}} p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) p_{\mathbf{B}_i}(\mathbf{b}_i) d\mathbf{b}_i\} \\
= & \sum_{i=1}^{m} \frac{\frac{\partial}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \int_{\mathcal{B}} p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) p_{\mathbf{B}_i}(\mathbf{b}_i) d\mathbf{b}_i}{\int_{\mathcal{B}} p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) p_{\mathbf{B}_i}(\mathbf{b}_i) d\mathbf{b}_i} \\
= & \sum_{i=1}^{m} \frac{\{E_{\mathbf{B}_i|\mathbf{Y}_i} \{\frac{\partial \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T}\}\} p_{\mathbf{Y}_i}(\mathbf{Y}_i)}{p_{\mathbf{Y}_i}(\mathbf{Y}_i)} \\
= & \sum_{i=1}^{m} E_{\mathbf{B}_i|\mathbf{Y}_i} \{\frac{\partial \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T}\}.
\end{aligned}
$$

The second derivative of marginal likelihood is

$$
\begin{aligned}
& \frac{\partial^2 l(\zeta, \mathbf{c}, \mathbf{d})}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T \partial(\mathbf{c}^T, \mathbf{d}^T)} \\
= & \sum_{i=1}^{m} E_{\mathbf{B}_i|\mathbf{Y}_i}(G_i) \\
= & \frac{\partial}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \sum_{i=1}^{m} \int_{\mathcal{B}} \frac{\partial \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \frac{p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) p_{\mathbf{B}_i}(\mathbf{b}_i)}{p_{\mathbf{Y}_i}(\mathbf{Y}_i)} d\mathbf{b}_i \\
= & \sum_{i=1}^{m} \{ \int_{\mathcal{B}} \frac{\partial^2 \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T \partial(\mathbf{c}^T, \mathbf{d}^T)} \frac{p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) p_{\mathbf{B}_i}(\mathbf{b}_i)}{p_{\mathbf{Y}_i}(\mathbf{Y}_i)} d\mathbf{b}_i + \\
& \int_{\mathcal{B}} \frac{\partial \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \frac{\partial}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \frac{p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) p_{\mathbf{B}_i}(\mathbf{b}_i)}{p_{\mathbf{Y}_i}(\mathbf{Y}_i)} d\mathbf{b}_i \} \\
= & \sum_{i=1}^{m} \int_{\mathcal{B}} \frac{\partial^2 \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T \partial(\mathbf{c}^T, \mathbf{d}^T)} \frac{p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) p_{\mathbf{B}_i}(\mathbf{b}_i)}{p_{\mathbf{Y}_i}(\mathbf{Y}_i)} d\mathbf{b}_i \\
& + \sum_{i=1}^{m} \int_{\mathcal{B}} \{ \frac{\partial \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \}^2 \frac{p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) p_{\mathbf{B}_i}(\mathbf{b}_i)}{p_{\mathbf{Y}_i}(\mathbf{Y}_i)} d\mathbf{b}_i \\
& - \sum_{i=1}^{m} \{ \int_{\mathcal{B}} \frac{\partial \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T} \frac{p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i) p_{\mathbf{B}_i}(\mathbf{b}_i)}{p_{\mathbf{y}_i}(\mathbf{Y}_i)} d\mathbf{b} \}^2 \\
= & \sum_{i=1}^{m} \{ E_{\mathbf{B}_i|\mathbf{Y}_i}(\frac{\partial^2 \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T \partial(\mathbf{c}^T, \mathbf{d}^T)}) + E_{\mathbf{B}_i|\mathbf{y}_i}(\frac{\partial \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T})^2 \\
& - [E_{\mathbf{B}_i|\mathbf{Y}_i}(\frac{\partial \log[p_{\mathbf{Y}_i|\mathbf{B}_i}(\mathbf{Y}_i)]}{\partial(\mathbf{c}^T, \mathbf{d}^T)^T})]^2 \}.
\end{aligned}
$$

Appendix B

Quadratic Approximation

In this section, we use the quadratic approximation to approximate the log marginal likelihood (4.2) at $\widetilde{g}$. We will simply use $p(\mathbf{b})$ for the probability density of $\mathbf{B}$. The log marginal likelihood for the subject ω_i is

$$l_i = \log \int_{\mathcal{B}} \frac{\exp\{\sum_{j=1}^{n_i}[g(Y_{ij}, X_{ij}) + b_i(Y_{ij}, X_{ij})]\}}{\prod_{j=1}^{n_i}[\int_{\mathcal{Y}} \exp\{g(y, X_{ij}) + b_i(y, X_{ij})\}dy]} p(\mathbf{b}_i)d\mathbf{b}_i.$$

We start from approximating l_i. Set

$$\begin{aligned} & L_{f,g}(\alpha) \\ = & \log \int_{\mathcal{B}} \frac{e^{\sum(f+\alpha g+b)}}{\prod[\int_{\mathcal{Y}} e^{f+\alpha g+b}dy]} p(\mathbf{b})d\mathbf{b} \\ = & \log \int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}, \end{aligned}$$

where

$$\begin{aligned} A(\alpha) &= \sum(f + \alpha g + b), \\ J(\alpha) &= \sum \log \int_{\mathcal{Y}} e^{f+\alpha g+b}dy, \\ h &= \log p(\mathbf{b}). \end{aligned}$$

Then,

$$\begin{aligned}
&\log \int_{\mathcal{B}} \frac{e^{\sum(g+b)}}{\prod(\int_{\mathcal{Y}} e^{g+b}dy)} p(\mathbf{b})d\mathbf{b} \\
= \;& L_{\widetilde{g},g-\widetilde{g}}(1) && \text{(B.1)} \\
\approx \;& L_{\widetilde{g},g-\widetilde{g}}(0) + L'_{\widetilde{g},g-\widetilde{g}}(0) + \frac{1}{2}L''_{\widetilde{g},g-\widetilde{g}}(0), && \text{(B.2)}
\end{aligned}$$

where $L^{(m)}_{\widetilde{g},g-\widetilde{g}}(0) = \frac{d^m}{d\alpha^m} L_{\widetilde{g},g-\widetilde{g}}(\alpha)|_{\alpha=0}$. We need $L'_{f,g}(0)$, $L''_{f,g}(0)$ for the approximation.

The first derivatives are

$$L'_{f,g}(\alpha) = \frac{\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}}{\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}},$$

$$\begin{aligned}
& A'(\alpha) \\
= \;& \frac{d}{d\alpha}[\sum(f+\alpha g) + \sum b] \\
= \;& \sum g, \\
A'(0) \;=\; & \sum g,
\end{aligned}$$

$$\begin{aligned}
& J'(\alpha) \\
= \;& \frac{d}{d\alpha}[\sum \log \int e^{f+\alpha g+b}dy] \\
= \;& \sum \frac{\int_{\mathcal{Y}} g e^{f+\alpha g+b}dy}{\int_{\mathcal{Y}} e^{f+\alpha g+b}dy}, \\
& J'(0) \\
= \;& \sum \frac{\int_{\mathcal{Y}} g e^{f+b}dy}{\int_{\mathcal{Y}} e^{f+b}dy} \\
\triangleq \;& \sum \mu_f(g|\mathbf{b}),
\end{aligned}$$

$$L'_{f,g}(0) = E^f_{\mathbf{B}|\mathbf{Y}}[\sum g - \sum \mu_f(g|\mathbf{B})].$$

The second derivatives are

$$\begin{aligned}
& L''_{f,g}(\alpha) \\
= & \frac{[\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}]' \int_{\mathcal{B}} e^{A-J+h}d\mathbf{b} - [\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}]' \int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}}{(\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b})^2} \\
= & \frac{[\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}]' \int_{\mathcal{B}} e^{A-J+h}d\mathbf{b} - [\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}]^2}{(\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b})^2} \\
= & \frac{[\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}]'}{\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}} - \{\frac{\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}}{\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}}\}^2,
\end{aligned}$$

where

$$[\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}]' = \int_{\mathcal{B}} e^{A-J+h}(A'-J')^2 - (A''-J'')e^{A-J+h}d\mathbf{b}$$

and

$$[\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}]' = \int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}.$$

since $A''(\alpha) = 0$, we have $A''(0) = 0$. In addition,

$$\begin{aligned}
& J''(\alpha) \\
= & \frac{d}{d\alpha}(\sum \frac{\int_{\mathcal{Y}} ge^{f+\alpha g+b}dy}{\int_{\mathcal{Y}} e^{f+\alpha g+b}dy}) \\
= & \sum \frac{(\int_{\mathcal{Y}} ge^{f+\alpha g+b}dy)' \int_{\mathcal{Y}} e^{f+\alpha g+b}dy - (\int_{\mathcal{Y}} e^{f+\alpha g+b}dy)' \int_{\mathcal{Y}} ge^{f+\alpha g+b}dy}{(\int_{\mathcal{Y}} e^{f+\alpha g+b}dy)^2} \\
= & \sum [\frac{\int_{\mathcal{Y}} g^2 e^{f+\alpha g+b}dy}{\int_{\mathcal{Y}} e^{f+\alpha g+b}dy} - (\frac{\int_{\mathcal{Y}} ge^{f+\alpha g+b}dy}{\int_{\mathcal{Y}} e^{f+\alpha g+b}dy})^2], \\
& J''(0) \\
= & \sum [\frac{\int_{\mathcal{Y}} g^2 e^{f+b}dy}{\int_{\mathcal{Y}} e^{f+b}dy} - (\frac{\int_{\mathcal{Y}} ge^{f+b}dy}{\int_{\mathcal{Y}} e^{f+b}dy})^2] \\
\triangleq & \sum V(g|\mathbf{b}).
\end{aligned}$$

Hence,

$$
\begin{aligned}
& L''_{f,g}(\alpha) \\
= \; & \frac{[\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}]'}{\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}} - \{\frac{\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}}{\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}}\}^2 \\
= \; & \frac{\int_{\mathcal{B}} e^{A-J+h}(A'-J')^2 - (A''-J'')e^{A-J+h}d\mathbf{b}}{\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}} - \{\frac{\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}}{\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}}\}^2 \\
= \; & \frac{\int_{\mathcal{B}} e^{A-J+h}(A'-J')^2 d\mathbf{b}}{\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}} - \frac{\int_{\mathcal{B}} (A''-J'')e^{A-J+h}d\mathbf{b}}{\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}} - \{\frac{\int_{\mathcal{B}} e^{A-J+h}(A'-J')d\mathbf{b}}{\int_{\mathcal{B}} e^{A-J+h}d\mathbf{b}}\}^2,
\end{aligned}
$$

$$
\begin{aligned}
& L''_{f,g}(0) \\
= \; & E^f_{\mathbf{B}|\mathbf{Y}}[\sum g - \mu_f(g|\mathbf{B})]^2 - E^f_{\mathbf{B}|\mathbf{Y}}[-nV(g|\mathbf{B})] - \{E^f_{\mathbf{B}|\mathbf{Y}}[\sum g - \mu_f(g|\mathbf{B})]\}^2 \\
= \; & E^f_{\mathbf{B}|\mathbf{Y}}[nV(g|\mathbf{B})] + V^f_{\mathbf{B}|\mathbf{Y}}[\sum g - \mu_f(g|\mathbf{B})] \\
= \; & E^f_{\mathbf{B}|\mathbf{Y}}[nV(g|\mathbf{B})] + V^f_{\mathbf{B}|\mathbf{Y}}[\mu_f(g|\mathbf{B})].
\end{aligned}
$$

We now put pieces together. Since

$$L_{\widetilde{g},g-\widetilde{g}}(\alpha) \approx L_{\widetilde{g},g-\widetilde{g}}(0) + L'_{\widetilde{g},g-\widetilde{g}}(0)\alpha + \frac{1}{2}L''_{\widetilde{g},g-\widetilde{g}}(0)\alpha^2.$$

And we have

$$
\begin{aligned}
L_{\widetilde{g},g-\widetilde{g}}(0) &= \log \int_{\mathcal{B}} \frac{e^{\sum(\widetilde{g}+b)}}{\prod(\int_{\mathcal{Y}} e^{\widetilde{g}+b}dy)} p(\mathbf{b})d\mathbf{b}, && \text{(B.3)} \\
L'_{\widetilde{g},g-\widetilde{g}}(0) &= E^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\sum(g-\widetilde{g}) - \sum \mu_{\widetilde{g}}(g-\widetilde{g}|\mathbf{B})], && \text{(B.4)} \\
L''_{\widetilde{g},g-\widetilde{g}}(0) &= \{E^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\sum V_{\widetilde{g}}(g-\widetilde{g}|\mathbf{B})] + V^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\mu_{\widetilde{g}}(g-\widetilde{g}|\mathbf{B})]\}. && \text{(B.5)}
\end{aligned}
$$

Plug (B.2), (B.3) and (B.4) in (B.1) , we have

$$
\begin{aligned}
& \log \int_{\mathcal{B}} \frac{e^{\sum g+b}}{(\int_{\mathcal{Y}} e^{g+b}dy)^n} p(\mathbf{b})d\mathbf{b} \\
\approx \; & \log \int_{\mathcal{B}} \frac{e^{\sum(\widetilde{g}+b)}}{\prod(\int_{\mathcal{Y}} e^{\widetilde{g}+b}dy)} p(\mathbf{b})d\mathbf{b} + \sum(g-\widetilde{g}) - E^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\sum \mu_{\widetilde{g}}(g-\widetilde{g}|\mathbf{B})] \\
& + \frac{1}{2}\{E^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\sum V_{\widetilde{g}}(g-\widetilde{g}|\mathbf{B})] + V^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\mu_{\widetilde{g}}(g-\widetilde{g}|\mathbf{B})]\}.
\end{aligned}
$$

Drop the term $V^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\mu_{\widetilde{g}}(g-\widetilde{g}|\mathbf{B})]$ for computational stability and terms do not involve g, we have

$$\begin{aligned}
&\log \int_{\mathcal{B}} \frac{e^{\sum g+b}}{(\int_{\mathcal{Y}} e^{g+b}dy)^n} p(\mathbf{b})d\mathbf{b} \\
\approx\ & \sum g - E^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\sum \mu_{\widetilde{g}}(g|\mathbf{B})] - E^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\sum V_{\hat{g}}(g,\widetilde{g}|\mathbf{B})] \qquad \text{(B.6)} \\
& + \frac{1}{2} E^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\sum V_{\widetilde{g}}(g|\mathbf{B})]. \qquad \text{(B.7)}
\end{aligned}$$

Define

$$L_{\widetilde{g}} = \mu_{\widetilde{g}}(g) - V_{\widetilde{g}}(\widetilde{g},g) + \frac{1}{2}V_{\widetilde{g}}(g,g),$$

where

$$\begin{aligned}
\mu_{\widetilde{g}}(g) &= \frac{1}{N}\sum_{i=1}^{m} E^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\sum_{j=1}^{n_i} \mu_{\widetilde{g}}(g|\mathbf{B})], \\
V_{\widetilde{g}}(\widetilde{g},g) &= \frac{1}{N}\sum_{i=1}^{m} E^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\sum_{j=1}^{n_i} V_{\widetilde{g}}(g,\widetilde{g}|\mathbf{B})], \\
V_{\widetilde{g}}(g,g) &= \frac{1}{N}\sum_{i=1}^{m} E^{\widetilde{g}}_{\mathbf{B}|\mathbf{Y}}[\sum_{j=1}^{n_i} V_{\widetilde{g}}(g|\mathbf{B})].
\end{aligned}$$

Therefore the quadratic approximation to log marginal likelihood (4.2) at $\widetilde{g}$ is

$$-\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_i} g(Y_{ij},X_{ij}) + L_{\widetilde{g}}.$$

Bibliography

Aubin, J. and Leoni-Aubin, S. (2008). Projection density estimation under a m-sample semiparametric model, *Computational Statistics and Data Analysis* **52**: 2451–2468.

Barndorff-Nielsen, O. and SrensenSource, M. (1994). A review of some aspects of asymptotic likelihood theory for stochastic processes, *International Statistical Review / Revue Internationale de Statistique* **62**: 133–165.

Booth, J. and Hobert, J. (1999). Maximizing generalized linear mixed model likelihoods with an automated monte carlo em algorithm, *Journal of the Royal Statistical Society B* **61**: 265–285.

Breunig, R. (2001). Density estimation for clustered data, *Econometric Reviews* **20**: 353–367.

Casella, G. and Berger, R. (2002). *Statistical Inference*, Duxbury Press, New York.

Christensen, R., Johnson, W., Branscum, A. and Hanson, T. (2011). *Bayesian Ideas and Data Analysis*, Chapman and Hall, London.

Davidow, J. H., Bothe, A., Andreatta, R. and Ye, J. (2009). Measurement of phonated intervals during four fluency-inducing conditions, *Journal of Speech, Language, and Hearing Research* **52**: 188–205.

Davison, A. and Hinkley, D. (1997). *Bootstrap Methods and their applications*, Cambridge University Press, Cambridge, UK.

Dias, R. (1998). Density estimation via hybrid splines, *Journal of Statistical Computation and Simulation* **60**: 277–294.

Dunson, D. (2007). Empirical bayes density regression, *Statistica Sinica* **17**: 481–504.

Dunson, D. B., Pillai, N. S. and Park, J. (2007). Bayesian density regression, *Journal of the Royal Statistical Society B* **69**: 163–183.

Efromovich, S. (2007). Conditional density estimation in a regression setting, *The Annals of Statistics* **35**: 2504–2535.

Efron, B. and Tibshirani, R. (1996). Using specially designed exponential families for density estimation, *The Annals of Statistics* **24**: 2431–2461.

Fan, J. and Yim, T. H. (2004). A crossvalidation method for estimating conditional densities, *Biometrika* **91**: 819–834.

Gehringer, K. R. and Redner, R. A. (1992). Nonparametric density estimation using normalized b-splines, *Communications in Statistics - Simulation and Computation* **21**: 849–878.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2003). *Bayesian Data Analysis*, Chapman and Hall, London.

Ghosh, K. and Jammalamadaka, S. (2001). Ageneral estimation method using spacings, *Journal of Statistical Planning and Inference* **93**: 71–82.

Givens, G. and Hoeting, J. (2005). *Computational Statistics*, John Wiley and Sons, New York.

Godinho, T., Ingham, R., Davidow, J. and Cotton, J. (2006). The distribution of phonated intervals in the speech of individuals who stutter, *Journal of Speech, Language and Hearing Research* **49**: 161–171.

Gow, M. and Ingham, R. J. (1992). Modifying electroglottogrph-identified intervals of phonation: The effect on stuttering, *Journal of speech, Language, and Hear Research* **35**: 495–511.

Griffin, J., Kolossiatis, M. and Steel, M. (2013). Comparing distributions by using dependent normalized random-measure mixtures, *Journal of the Royal Statistical Society B* **75**: 499–529.

Gu, C. (1993). Smoothing spline density estimation: A dimensionless automatic algorithm, *Journal of the American Statistical Association* **88**: 495–504.

Gu, C. (1995). Smoothing spline density estimation: Conditional distribution, *Statistica Sinica* **5**: 709–736.

Gu, C. (2009). General smoothing spline, Available at http://cran.r-project.org.

Gu, C. (2013). *Smoothing Spline ANOVA Models*, Springer-Verlag, NewYork.

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory, *The Annals of Statistics* **21**: 217–234.

Gu, C. and Wahba, G. (1991). Comments to 'multivariate adaptive regression splines', by j. friedman, *The Annals of Statistics* **19**: 115–123.

Gu, C. and Wahba, G. (1993). Semiparametric analysis of variance with tensor product thin plate splines, *Journal of the Royal Statistical Society B* **55**: 353–368.

Gu, C. and Wang, J. (2003). Penalized likelihood density estimation:direct cross-validation and scalable approximation, *Statistica Sinica* **13**: 811–826.

Gu, M. G. and Kong, F. H. (1998). A stochastic approximation algorithm with markov chain monte-carlo for incomplete data estimation problem, *Proceedings of the National Academy of Sciences* **95**: 7270–7274.

Gu, M. G. and Kong, F. H. (2000). A generalized markov chain monte carlo stochastic approximation algorithm for statistical computing. Personal communication.

Gu, M. G. and Zhu, H. T. (2001). Maximum likelihood estimation for spatial models by markov chain monte carlo stochastic approximation, *Journal of the Royal Statistical Society B* **63**: 339–355.

Hall, P., Lahiri, S. N. and Truong, Y. K. (1995). On bandwidth choice for density estimation with dependent data, *The Annals of Statistics* **23**: 2241–2263.

Hall, P., Racine, J. and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities, *Journal of the American Statistical Association* **99**: 1015–1026.

Hart, J. D. and Vieu, P. (1990). Data-driven bandwidth choice for density estimation based on dependent data, *Annals of Statistics* **18**: 873–890.

Ingham, R. J., Kilgo, M., Ingham, J., Moglia, R., Belknap, H. and Sanchez, T. (2001). Evaluation of a stuttering treatment based on reduction of short phonation intervals, *Journal of Speech,Language, and Hear Research* **44**: 1229–1244.

Jennrich, R. I. and Schluchter, M. D. (1986). Unbalanced repeated-measures model with structured covariance matrices, *Biometrics* **42**: 805–820.

Jiang, Y., Karcher, P. and Wang, Y. (2011). On implementation of the markov chain monte carlo stochastic approximation algorithm, *Advances in Directional and Linear Statistics: A Festchrift for Srenivasa Rao Jammalamadaka* pp. 97–111. In M.T. Wells and A. Sengupta (eds.).

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise, *Journal of the Royal Statistical Society B* **59**: 319–351.

Karcher, P. and Wang, Y. (2001). Generalized nonparametric mixed effects models, *Journal of Computational and Graphical Statistics* **10**: 641–655.

Ke, C. and Wang, Y. (2001). Semi-parametric nonlinear mixed effects models and their applications (with discussion), *Journal of the American Statistical Association* **96**: 1272–1298.

Kooperberg, C. and Stone, C. (1991). A study of logspline density estimation, *Computational Statistics and Data Analysis* **12**: 327–347.

Lai, T. L. (2003). Stochastic approximation, *Annals of Statistics* **31**: 391–406.

Lenk, P. (1988). The logistic normal distribution for bayesian, nonparametric, predictive densities, *Journal of the American Statistical Association* **83**: 509–516.

Lenk, P. (1998). Towards a practicable bayesian nonparametric density estimator, *Biometrika* **78**: 531–543.

Leonard, T. (1978). Density estimation, stochastic processes, and prior information, *Journal of the Royal Statistical Society B* **40**: 113–146.

O'Sullivan, F. (1998). Fast computation of fully automated log-density and log-hazard estimators, *SIAM Journal on Scientific and Statistical Computing* **9**: 363–379.

Parzen, E. (1962). On estimation of a probability density function and mode, *Annals of Mathematical Statistics* **33**: 1065–1076.

Qin, J. and Zhang, B. (2005). Density estimation under a two-sample semiparametric model, *Nonparametric Statistics* **17**: 665–683.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves, *Journal of the Royal Statistical Society B* **53**: 233–243.

Robbins, H. and Monro, S. (1951). A stochastic approximation method, *The Annals of Mathematical Statistics* **22**: 400–407.

Rodriguez, A. and Ter Horst, E. (2008). Bayesian dynamic density estimation, *Bayesian Analysis* **3**: 339–366.

Rodriguez, A., Dunson, D. B. and Taylor, J. (2009). Bayesian hierarchically weighted finite mixtures models for samples of distributions, *Biostatistics* **10**: 155–171.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

Stone, C. (1990). Large sample inference for logspline model, *Annals of Statistics* **18**: 717–741.

Striebel, T. (1959). Densities for stochastic processes, *The Annals of Mathematical Statistics* **30**: 559–567.

Wahba, G. (1981). Data-based optimal smoothing of orthogonal series density estimates, *Annals of Statistics* **9**: 146–156.

Wahba, G. (1990). *Spline models for observational data*, Vol. 59, Regional Conference Series in Applied Mathematics.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. (1995). Smoothing spline anova for exponential families with application to the wisconsin epidemiological study of diabetic retinopathy, *Annals of Statistics* **23**: 1865–1895.

Wang, Y. (1998). Mixed effects smoothing spline analysis of variance, *Journal of the Royal Statistical Society B* **60**: 159–174.

Wang, Y. (2011). *Smoothing Splines: Methods and Applications*, Chapman and Hall, New York.